Speculations Concerning the First Ultraintelligent Machine*

IRVING JOHN GOOD

Trinity College, Oxford, England and
Atlas Computer Laboratory, Chilton, Berkshire, England

1. Introduction

The survival of man depends on the early construction of an ultraintelligent machine.

In order to design an ultraintelligent machine we need to understand more about the human brain or human thought or both. In the following pages an attempt is made to take more of the magic out of the brain by means of a "subassembly" theory, which is a modification of Hebb's famous speculative cell-assembly theory. My belief is that the first ultraintelligent machine is most likely to incorporate vast artificial neural circuitry, and that its behavior will be partly explicable in terms of the subassembly theory. Later machines will all be designed by ultra-

* Based on talks given in a Conference on the Conceptual Aspects of Biocommunications, Neuropsychiatric Institute, University of California, Los Angeles, October 1962; and in the Artificial Intelligence Sessions of the Winter General Meetings of the IEEE, January 1963 [*1, 46*].

The first draft of this monograph was completed in April 1963, and the present slightly amended version in May 1964.

I am much indebted to Mrs. Euthie Anthony of IDA for the arduous task of typing.

intelligent machines, and who am I to guess what principles they will devise? But probably Man will construct the *deus ex machina* in his own image.

The subassembly theory sheds light on the physical embodiment of memory and meaning, and there can be little doubt that both will need embodiment in an ultraintelligent machine. Even for the brain, we shall argue that physical embodiment of meaning must have originated for reasons of economy, at least if the metaphysical reasons can be ignored. Economy is important in any engineering venture, but especially so when the price is exceedingly high, as it most likely will be for the first ultraintelligent machine. Hence semantics is relevant to the design of such a machine. Yet a detailed knowledge of semantics might not be required, since the artificial neural network will largely take care of it, provided that the parameters are correctly chosen, and provided that the network is adequately integrated with its sensorium and motorium (input and output). For, if these conditions are met, the machine will be able to learn from experience, by means of positive and negative reinforcement, and the instruction of the machine will resemble that of a child. Hence it will be useful if the instructor knows something about semantics, but not necessarily more useful than for the instructor of a child. The correct choice of the parameters, and even of the design philosophy, will depend on the usual scientific method of successive approximation, using speculation, theory, and experiment. The percentage of speculation needs to be highest in the early stages of any endeavor. Therefore no apology is offered for the speculative nature of the present work. For we are certainly still in the early stages in the design of an ultraintelligent machine.

In order that the arguments should be reasonably self-contained, it is necessary to discuss a variety of topics. We shall define an ultraintelligent machine, and, since its cost will be very large, briefly consider its potential value. We say something about the physical embodiment of a word or statement, and defend the idea that the function of meaning is economy by describing it as a process of "regeneration." In order to explain what this means, we devote a few pages to the nature of communication. (The brain is of course a complex communication and control system.) We shall need to discuss the process of recall, partly because its understanding is very closely related to the understanding of understanding. The process of recall in its turn is a special case of statistical information retrieval. This subject will be discussed in Section 5. One of the main difficulties in this subject is how to estimate the probabilities of events that have never occurred. That such probabilities are relevant to intelligence is to be expected, since intelligence is sometimes defined as the ability to adapt to new circumstances.

The difficulty of estimating probabilities is sometimes overlooked in the literature of artificial intelligence, but this article would be too long if the subject were surveyed here. A separate monograph has been written on this subject [*48*].

Some of the ideas of Section 5 are adapted, in Section 6, to the problem of recall, which is discussed and to some extent explained in terms of the subassembly theory.

The paper concludes with some brief suggestions concerning the physical representation of "meaning."

This paper will, as we said, be speculative: no blueprint will be suggested for the construction of an ultraintelligent machine, and there will be no reference to transisitors, diodes, and cryogenics. (Note, however, that cryogenics have the important merit of low power consumption. This feature will be valuable in an ultraintelligent machine.) One of our aims is to pinpoint some of the difficulties. The machine will not be on the drawing board until many people have talked big, and others have built small, conceivably using deoxyribonucleic acid (DNA).

Throughout the paper there are suggestions for new research. Some further summarizing remarks are to be found in the Conclusions.

2. Ultraintelligent Machines and Their Value

Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind (see for example refs. [*22*], [*34*], [*44*]). Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.

In one science fiction story a machine refused to design a better one since it did not wish to be put out of a job. This would not be an insuperable difficulty, even if machines can be egotistical, since the machine could gradually improve *itself* out of all recognition, by acquiring new equipment.

B. V. Bowden stated on British television (August 1962) that there is no point in building a machine with the intelligence of a man, since it is easier to construct human brains by the usual method. A similar point was made by a speaker during the meetings reported in a recent IEEE

publication [*1*], but I do not know whether this point appeared in the published report. This shows that highly intelligent people can overlook the "intelligence explosion." It is true that it would be uneconomical to build a machine capable *only* of ordinary intellectual attainments, but it seems fairly probable that if this could be done then, at double the cost, the machine could exhibit ultraintelligence.

Since we are concerned with the economical construction of an ultra-intelligent machine, it is necessary to consider first what such a machine would be worth. Carter [*11*] estimated the value, to the world, of J. M. Keynes, as at least 100,000 million pounds sterling. By definition, an ultraintelligent machine is worth far more, although the sign is uncertain, but since it will give the human race a good chance of surviving indefinitely, it might not be extravagant to put the value at a mega-keynes. There is the opposite possibility, that the human race will become redundant, and there are other ethical problems, such as whether a machine could feel pain especially if it contains chemical artificial neurons, and whether an ultraintelligent machine should be dismantled when it becomes obsolete [*43, 84*]. The machines will create social problems, but they might also be able to solve them in addition to those that have been created by microbes and men. Such machines will be feared and respected, and perhaps even loved. These remarks might appear fanciful to some readers, but to the writer they seem very real and urgent, and worthy of emphasis outside of science fiction.

If we could raise say a hundred billion dollars we might be able to simulate all the neurons of a brain, and of a whole man, at a cost of ten dollars per artificial neuron. But it seems unlikely that more than say a millikeynes would actually be forthcoming, and even this amount might be difficult to obtain without first building the machine! It would be justified if, with this expenditure, the chance of success were about 10^{-9}.

Until an ultraintelligent machine is built perhaps the best intellectual feats will be performed by men and machines in very close, sometimes called "symbiotic," relationship, although the term "biomechanical" would be more appropriate. As M. H. A. Newman said in a private communication in 1946, an electronic computer might be used as "rough paper" by a mathematician. It could already be used in this manner by a chess player quite effectively, although the effectiveness would be much increased if the chess-playing programs were written with extremely close man-machine interaction in mind from the start. The reason for this effectiveness is that the machine has the advantage in speed and accuracy for routine calculation, and man has the advantage in imagination. Moreover, a large part of imagination in chess can be reduced to routine. Many of the ideas that require imagination in the amateur are routine for the master. Consequently the machine might

appear imaginative to many observers and even to the programmer. Similar comments apply to other thought processes.

The justification for chess-playing programs is that they shed light on the problem of artificial intelligence without being too difficult to write. Their interest would be increased if chess were replaced by so-called "randomized chess," in which the positions of the white pieces on the first rank are permuted at random before the game begins (but with the two bishops on squares of opposite colors), and then the initial positions of the black pieces are determined by mirror symmetry. This gives rise to 1440 essentially distinct initial positions and effectively removes from the game the effect of mere parrot learning of the openings, while not changing any of the general principles of chess. In ordinary chess the machine would sometimes beat an international Grandmaster merely by means of a stored opening trap, and this would be a hollow victory. Furthermore a program for randomized chess would have the advantage that it would not be necessary to store a great number of opening variations on magnetic tape.

The feats performed by very close man-machine interaction by say 1980 are likely to encourage the donation of large grants for further development. By that time, there will have been great advances in microminiaturization, and pulse repetition frequencies of one billion pulses per second will surely have been attained in large computers (for example see Shoulders [*91*]). On the other hand, the cerebral cortex of a man has about five billion neurons, each with between about twenty and eighty dendrites ([*90*], pp. 35 and 51), and thousands of synapses. (At the recent IEEE meetings, P. Mueller offered the estimate 300,000 orally. It would be very interesting to know the corresponding figure for the brain of a whale, which, according to Tower [*99*], has about three times as many neurons as a human brain. Perhaps some whales are ultraintelligent! [*49*].) Moreover, the brain is a parallel-working device to an extent out of all proportion to any existing computer. Although computers are likely to attain a pulse repetition speed advantage of say a million over the brain, it seems fairly probable, on the basis of this quantitative argument, that an ultraintelligent machine will need to be ultraparallel.

In order to achieve the requisite degree of ultraparallel working *it might be useful for many of the elements of the machine to contain a very short-range microminiature radio transmitter and receiver*. The range should be small compared with the dimensions of the whole machine. A "connection" between two close artificial neurons could be made by having their transmitter and receiver on the same or close frequencies. The strength of the connection could be represented by the accuracy of the tuning. The receivers would need numerous filters so as to be

capable of receiving on many different frequencies. "Positive reinforcement" would correspond to improved tuning of these filters.

It cannot be regarded as entirely certain that an ultraintelligent machine would need to be ultraparallel, since the number of binary operations per second performed by the brain might be far greater than is necessary for a computer made of reliable components. Neurons are not fully reliable; for example, they do not all last a lifetime; yet the brain is extremely efficient. This efficiency must depend partly on "redundancy" in the sense in which the term is used in information theory. A machine made of reliable components would have an advantage, and it seems *just* possible that ultraparallel working will not be essential. But there is a great waste in having only a small proportion of the components of a machine active at any one time.

Whether a machine of classical or ultraparallel design is to be the first ultraintelligent machine, it will need to be able to handle or to learn to handle ordinary language with great facility. This will be important in order that its instructor should be able to teach it rapidly, and so that later the machine will be able to teach the instructor rapidly. It is very possible also that natural languages, or something analogous to them rather than to formal logic, are an essential ingredient of scientific imagination. Also the machine will be called upon to translate languages, and perhaps to generate fine prose and poetry at high speed, so that, all in all, linguistic facility is at a high premium.

A man cannot learn more than ten million statements in a lifetime. A machine could already *store* this amount of information without much difficulty, even if it were not ultraparallel, but it seems likely that it would need to be ultraparallel in order to be able to retrieve the information with facility. It is in recall rather than in retention that the ordinary human memory reveals its near magic. The greatest mental achievements depend on more than memory, but it would be a big step toward ultraintelligence if human methods of recall could be simulated.

For the above reasons, it will be assumed here that the first ultraintelligent machine will be ultraparallel, perhaps by making use of radio, as suggested. For definiteness, the machine will be assumed to incorporate an artificial neural net. This might be in exceedingly close relationship with an ordinary electronic computer, the latter being used for the more formalizable operations [*33*]. In any event the ultraintelligent machine might as well have a large electronic computer at its beck and call, and also a multimillion dollar information-retrieval installation of large capacity but of comparatively slow speed, since these would add little to the total cost.

It is unlikely that facility in the use of language will be possible if semantic questions are ignored in the design. When we have read or

listened to some exposition we sometimes remember for a long time what it meant, but seldom how it was worded. It will be argued below that, for men, meaning serves a function of economy in long-term retention and in information handling, and this is the basis for our contention that semantics are relevant to the design of an ultraintelligent machine.

Since language is an example of communication, and since an ultraintelligent machine will be largely a complicated communication system, we shall briefly consider the nature of communication. It will be argued that in communication a process of "generalized regeneration" always occurs, and that it serves a function of economy. It will also be argued that the meanings of statements are examples of generalized regeneration.

3. Communication as Regeneration[1]

In a communication system, a source, usually a time series denoted here by $S(t)$ or S for short, undergoes a sequence of transformations. The first transformation is often a deterministic *encoding*, which transforms the source into a new time series, $T_0S(t)$. This is noisily (indeterministically) transmitted, i.e., it undergoes a transformation T_1 which is a random member of some class of transformations. If the possible sources are, in some sense, far enough apart, and if the noise is not too great, then the waveforms T_1T_0S will, also in some sense, tend to form clumps, and it will be possible with high probability to reconstruct the encoded sources at the receiving end of the channel. This reconstruction is called here (generalized) *regeneration*, a term that is most familiar in connection with the reshaping of square pulses. When dealing with groups of consecutive pulses, the term *error correction* is more usual, especially when it is assumed that the individual pulses have themselves been first regenerated. Another way of saying that the source signals must be far enough apart is to say that they must have enough *redundancy*. In a complicated network, it is often convenient to regard signals as sources at numerous places in the network and not merely at the input to the network. The redundancy might then be represented, for example, by mere serial or parallel repetition.

A compromise between pure regeneration and the use of the whole garbled waveform $T_1T_0S(t)$ is *probabilistic regeneration*, in which the garbled waveform is replaced by the set of probabilities that it has arisen from various sources [*42*]. In probabilistic regeneration less information is thrown away than in pure regeneration, and the later data-handling costs more, but less than it would cost if there were no regeneration at all. The hierarchical use of probabilistic regeneration would add much flexibility to complicated communication networks.

[1] For a short survey of the nature of communication, see for example Pierce [*80a*].

An example of generalized and hierarchical regeneration is in the use of words in a language. A *word* in a spoken language could well be defined as a clump of short time series, that is, a *class* of time series having various properties in common. (The class depends on the speaker, the listener, and the context; and membership of the class is probabilistic since there are marginal cases.) If any sound (acoustic time series) belonging to the clump is heard, then the listener mentally regenerates the sound and replaces it by some representation of the *word*. He will tend to remember, or to write down, the word and not its precise sound, although if any other significant property of the sound is noticed it might also be remembered. The advantage of remembering the word rather than the precise sound is that there is then less to remember and a smaller amount of information handling to do.

This process of regeneration occurs to some extent at each of the levels of phonemes, words, sentences, and longer linguistic stretches, and even at the semantic level, and wherever it occurs it serves a function of economy. But the economy is not parsimonious: "redundancy" often remains in the coding in order that the encoded message should continue to have useful error-correcting features. The redundancy often decreases with the passage of time, perhaps leading eventually to the extinction of a memory.

That communication theory has a bearing on the philosophy of meaning has been suggested before (see for example, Weaver [*89*], pp. 114–117, and Lord Brain [*8*]). Note also the definition of the amount of subjective information in a proposition, as $-\log_2 p$ where p is the initial subjective probability that the proposition is true ([*21*], p. 75). This could also be described as subjective semantic information: when the probabilities are credibilities (logical probabilities) we obtain what might be called objective semantic information [*5*, *10*], the existence of which is, in my opinion, slightly more controversial. That subjective probability is just as basic as communication theory to problems of meaning and recognition, if not more so, is a necessary tenet for any of us who *define* reasoning as logic plus probability ([*21*], pp. 3 and 88; see also Colin Cherry [*12*], pp. 200 and 274, Woodward [*105*], and Tompkins [*98*]). The implication is that both the initial (prior) probabilities and the likelihoods or "weights of evidence" [*21*] should be taken into account in every practical inference by a rational man, and in fact nearly always are taken into account to some extent, at least implicitly, even by actual men. (In case this thesis should appear as obvious to some readers as it does to the writer, it should be mentioned that in 1950 very few statisticians appeared to accept the thesis; and even now they are in a minority.) There is conclusive experimental evidence that the recognition of words depends on the initial proba-

bilities [*94*]: a well-known method of deception when trying to sell a speech-synthesis system is to tell the listeners in advance what will be said on it, and thereby to make it easily intelligible when it is repeated. There is a similar effect in the perception of color [*9*].

The rational procedure in perception would be to estimate the final (a posteriori) probabilities by means of Bayes' theorem, and then perhaps to select one or more hypotheses for future consideration or action, by allowing also for the utilities. (Compare refs. [*24*], [*12*], p. 206, and Middleton [*68*].) In fact the "principle of rationality" has been defined as the recommendation to maximize expected utility. But it is necessary to allow also for the expected cost of information handling including theorizing [*23*, *40*], and this is why regeneration and probabilistic regeneration are useful.

We pointed out above that the organization of regeneration is often hierarchical, but it is not purely so. For example, we often delay the regeneration of a phoneme until the word to which the phoneme belongs has been regenerated with the help of the surrounding context. Likewise if a machine is to be able to "understand" ordinary spoken language in any reasonable sense, it seems certain that its corresponding regeneration structure must not be purely hierarchical unless it is also probabilistic. For each process of nought-one or pure regeneration (each definite "decision") loses information, and the total loss would certainly be too great unless the speech were enunciated with priggish accuracy. The probabilistic regeneration structure that will be required will be much more complex than a "pure" regeneration structure. (Historical note: the hierarchical structure of mental processes was emphasized by Gall [*20*], McDougall [*66*], and many times since—see for example MacKay [*63*], Hayek [*53*], and others [*30*], [*34*], [*87*], [*41*].)

It seems reasonable to the present writer that probabilistic regeneration will, for most purposes, lose only a small amount of information, and that, rather than to use anything more elaborate, it is likely to be better to compromise between pure and probabilistic regeneration for most purposes.

The applications of regeneration in the present paper will be to assemblies, subassemblies, and *meaning*. When a person recalls a proposition he could be said to regenerate its meaning; when he understands a statement made by another person the term "transgeneration" would be more appropriate; and when he thinks of a new proposition, the process would be better called "generation," but we shall use the word "regeneration" to cover all three processes. For example, when listening to speech, the production of meaning can be regarded as the last regeneration stage in the hierarchy mentioned before, and it performs a function of economy just as all the other stages do. It is possible

that this has been frequently overlooked because meaning is associated with the metaphysical nature of consciousness, and one does not readily associate metaphysics with questions of economy. Perhaps there is nothing more important than metaphysics, but, for the construction of an artificial intelligence, it will be necessary to represent meaning in some physical form.

4. Some Representations of "Meaning" and Their Relevance to Intelligent Machines

Semantics is not relevant to all problems of mechanical language processing. Up to a point, mechanical translation can be performed by formal processes, such as dictionary look-up and some parsing. Many lexical ambiguities can be resolved statistically in terms of the context, and some as a consequence of the parsing. Sometimes one can go further by using an iterative process, in which the lexical ambiguities are resolved by the parsing, and the parsing in its turn requires the resolution of lexical ambiguities. But even with this iterative process it seems likely that perfect translation will depend on semantic questions [*14*, *89*]. Even if this is wrong, the design of an ultraintelligent machine will still be very likely to depend on semantics [*31*, *50*]. What then is meant by semantics?

When we ask for the meaning of a statement we are talking about language, and are using a metalanguage; and when we ask for the meaning of "meaning" we are using a metametalanguage, so it is not surprising that the question is difficult to answer. A recent survey chapter was entitled "The Unsolved Problem of Meaning" [*3*]. Here we shall touch on only a few aspects of the problem, some of which were not mentioned in that survey (see also Black [*7*]).

It is interesting to recall the thought-word-thing triangle of Charles Pierce and of Ogden and Richards. (See, for example Cherry [*12*], p. 110. Max Black ascribed a similar "triangle" to the chemist Lavoisier in a recent lecture.) It will help to emphasize the requirement for a physical embodiment of meaning if it is here pointed out that the triangle could be extended to a thought-word-thing-engram tetrahedron, where the fourth vertex represents the physical embodiment of the word in the brain, and will be assumed here usually to be a cell assembly.

Given a class of linguistic transformations that transform statements into equivalent statements, it would be plausible to represent the meaning of the statement, or the proposition expressed by the statement, by the class of all equivalent statements. (This would be analogous to a modified form of the Frege-Russell definition of a cardinal

integer, for example, "3" can be defined as the class of all classes "similar" to the class consisting of the words "Newton," "Gauss," and "Bardot.") The point of this representation is that it makes reference to linguistic operations alone, and not to the "outside world." It might therefore be appropriate for a reasoning machine that had few robotic properties. Unfortunately, linguistic transformations having a strictly transitive property are rare in languages. There are also other logical difficulties analogous to those in the Frege-Russell definition of a cardinal integer. Moreover, this representation of meaning would be excessively unwieldy for mechanical use.

Another possible representation depending on linguistic transformations is a *single representative* of the class of all equivalent statements. This is analogous to another "definition" or, better, "representation," of a cardinal integer (see for example Halmos [*51*], p. 99). This representation is certainly an improvement on the previous one. If this representation were to be used in the construction of an ultraintelligent machine, it would be necessary to invent a language in which each statement could be reduced to a canonical form. Such an achievement would go most of the way to the production of perfect mechanical translation of technical literature, as has often been recognized, and it would also be of fundamental importance for the foundations of intuitive or logical probability ([*21*], pp. 4 and 48). The design of such a "canonical language" would be extremely difficult, perhaps even logically impossible, or perhaps it would require an ultraintelligent machine to do it!

For human beings, meaning *is* concerned with the outside world or with an imaginary world, so that representations of meaning that are not entirely linguistic in content might be more useful for our purpose. The behaviorist regards a statement as a stimulus, and interprets its meaning in terms of the class of its effects (responses) in overt behavior. The realism of this approach was shown when "Jacobson . . . made the significant discovery that action potentials arise in muscles simultaneously with the meaning processes with which the activity of the muscle, if overtly carried out, would correspond" ([*3*], p. 567). Thus the behavioral interpretation of meaning might be relevant for the understanding of the behavior and education of people and robots, especially androids. But, for the *design* of ultraintelligent machines, the internal representation of meaning (inside the machine) can hardly be ignored, so that the behavioral interpretation is hardly enough.

So far we have been discussing the interpretation and representation of the meaning of a statement, but even the meaning of a word is much less tangible and clear-cut than is sometimes supposed. This fact was emphasized, for example, by the philosopher G. E. Moore. Later John

Wisdom (not J. O. Wisdom) emphasized that we call an object a cow if it has enough of the properties of a cow, with perhaps no single property being essential. The need to make this interpretation of meaning more quantitative and probabilistic has been emphasized in various places by the present writer, who has insisted that this "probabilistic definition" is of basic important for future elaborate information-retrieval systems [*29*, *35*, *31*, *43*, *41*]. "An object is said to belong to class C (such as the class of cows) if some function $f(p_1, p_2, \ldots, p_m)$ is positive, where the p's are the credibilities (logical probabilities) that the object has qualities $Q_1, Q_2, \ldots, Q_m$. These probabilities depend on further functions related to other qualities, on the whole more elementary, and so on. *A certain amount of circulatory is typical.* For example, a connected brown patch on the retina is more likely to be caused by the presence of a cow if it has four protuberances that look like biological legs than if it has six; but each protuberance is more likely to be a biological leg if it is connected to something that resembles a cow rather than a table. In view of the circularity in this interpretation of "definition," the stratification in the structure of the cerbral cortex can be only a first approximation to the truth" ([*41*], pp. 124–125; see also Hayek [*53*], p. 70). The slight confusion in this passage, between the definition of a cow and the recognition of one, was deliberate, and especially appropriate in an anthology of partly baked ideas. It can be resolved by drawing the distinction between a logical probability and a subjective probability (see for example [*36*]), and also the distinction between subjective and objective information that we made in the previous section.

If we abandon interpretations of meaning in terms of linguistic transformations, such as dictionary definitions, or, in the case of statements, the two interpretations mentioned before; and if also we do not regard the behavioral interpretations as sufficient, we shall be forced to consider interpretations in terms of internal workings. Since this article is written mainly on the assumption that an ultraintelligent machine will consist largely of an artificial neural net, we need in effect a neurophysiological representation of meaning. The behavioral interpretation will be relevant to the education of the machine, but not so much to its design. It does not require much imagination to appreciate that the probabilistic and iterative interpretation of the definition of a word, as described above, is liable to fit well into models of the central nervous system.

It has been difficult for the writer to decide how much neurophysiology should be discussed, and hopefully an appropriate balance is made in what follows between actual neurophysiology and the logic of artificial neural networks. The discussion will be based on the speculative

cell-assembly theory of Hebb [*54*] (see also [*53*] and [*71*]), or rather on a modification of it in which "subassemblies" are emphasized and a central control is assumed. If the present discussion contains inconsistencies, the present writer should be blamed. (For a very good survey of the relevant literature of neurophysiology and psychology, see Rosenblatt [*82*], pp. 9–78.)

5. Recall and Information Retrieval

Whatever might be the physical embodiment of meaning, it is certainly closely related to that of long-term recall. *Immediate* recall is not strongly related to semantics, at any rate for linguistic texts. In fact, experiments show that immediate and exact recall of sequences of up to fifty words is about as good for meaningless texts as it is for meaningful texts, provided that the meaningless ones are at least "fifth order" approximations to English, that is to say that the probability of each word, given the previous five, is high [*70*].

The process of recall is a special case of information retrieval, so that one would expect there to be a strong analogy between the recall of a memory and the retrieval of documents by means of index terms. An immediate recall is analogous to the trivial problem of the retrieval of a document that is already in our hands. The problem of the retrieval of documents that are not immediately to hand is logically a very different matter, and so it is not surprising that the processes of immediate and long-term recall should also differ greatly.

The problem of what the physical representation is for immediate recall is of course not trivial, but for the moment we wish to discuss long-term recall since it is more related to the subject of semantics.

The usual method for attacking the problem of document retrieval, when there are many documents (say several thousand), is to index each document by means of several index terms. We imagine a library customer, in need of some information, to list some index terms without assuming that he uses any syntax, at least for the present. In a simple retrieval system, the customer's index terms can be used to extract documents by means of a sort, as of punched cards. The process can be made more useful, not allowing for the work in its implementation, if the terms of the documents, and also those of the customer, are given various weights, serving in some degree the function of probabilities. We then have a weighted or statistical system of information retrieval.

One could conceive of a more complicated information-retrieval system in which each document had associated with it a set of resonating filters forming a circuit C. All documents would be queried *in parallel*: the "is-there-a-doctor-in-the-house" principle [*86*]. The

amount of energy generated in the circuit C would be fed back to a master control circuit. (In the brain, the corresponding control system might be the "centrencephalic system" [79].) Whichever circuit C fed back the maximum power, the corresponding document would be extracted first. If this document alone failed to satisfy the customer completely, then the circuit C would be provisionally disconnected, and the process repeated, and so on.

Ideally, this search would be probabilistic, in the sense that the documents would be retrieved in order of descending a posteriori probability, and the latter would be registered also. If these were $p_1, p_2, \ldots$, then the process would stop at the nth document, where there would be a threshold on n, and on $p_1 + p_2 + \ldots + p_n$. For example, the process might stop when $n = 10$, or when $p_1 + p_2 + \ldots + p_n > 0.95$, whichever occurred first. The thresholds would be parameters, depending on the importance of the search. (For the estimation of probabilities, see [48].)

When we wish to recall a memory, such as a person's name, we consciously or unconsciously use *clues*, analogous to index terms. These clues are analogous to weighted index terms, and it seems virtually certain that they lead to the retrieval of the appropriate memory by means of a parallel search, just as in the above hypothetical document-retrieval system. The speed of neural conduction is much too slow for a primarily serial search to be made. The search might very well be *partly* serial: the less familiar memories take longer to recall and require more effort. This might be because the physical embodiment of the less familiar memory requires a greater weight of clues before it will "resonate" strongly enough.

Further evidence that the search is, on the whole, more parallel than serial can be derived from Mandelbrot's explanation of the Zipf "law" of distribution of words [28]. The explanation requires that the effort of extracting the rth commonest word from memory is roughly proportional to $\log r$. This is reasonable for a parallel search, whereas the effort would be roughly proportional to r for a serial search.

When the clues do spark off the sought memory, this memory in its turn reminds us of other clues that we might have used in advance if we had thought of doing so. These "retrieved clues" often provide an enormous factor in favor of the hypothesis that the memory retrieved is the one that was sought: consequently we are often morally certain that the memory is the right one once it is recalled, even though its recall might have been very difficult. There is again a strong resemblance to document retrieval.

When we extract a wrong memory, it causes incorrect clues to come to mind, and these are liable to block the correct memory for a number of

seconds, or for longer if we panic. This is another reason why the less familiar memories take longer to recall.

When we wish to recall incidents from memory, pertaining to a particular subject, the method used is to bring to mind various relevant facts and images in the hope that they are relevant enough, numerous enough, vivid enough, independent enough, and specific enough to activate the appropriate memory. (If specificity is lacking, then the wrong memory is liable to be recalled.) There is a clear analogy with the probabilistic definition of a word and probabilistic recognition of an object quoted in Section 4. A corresponding method of information retrieval is to list index terms that are relevant enough, numerous enough, independent enough, and specific enough, and (if the process is not entirely mechanized) vivid enough. This attitude towards index terms leads to forms of probabilistic or statistical indexing, as suggested independently by the writer ([35], [31], p. 12) and by Maron and Kuhns [64] who treated the matter in more detail. The present writer regards subjective and logical probabilities as partially ordered only [21], but does not consider that the fact of only partial ordering is the main source of the difficulties in probabilistic indexing.

We have said enough to bring out the analogy between the process of recall and the techniques of document retrieval, and to indicate that, if it is possible to develop a comprehensive theory of either of these subjects, it should be a probabilistic theory. The need for a probabilistic theory is further brought out by means of a short discussion of what might be called "statistical semantics."

A complete discussion of statistical semantics would lean heavily on (i) the very intricate subject of non-statistical semantics, and on (ii) some statistical theory concerning language, without any deep discussion of semantic problems. But our purpose in this section is only to make clear that a complete treatment of statistical semantics would be somewhat more general than recall and document retrieval.

If we wish to teach a language to a baby who starts in a state of comparative ignorance, we simultaneously allow him to become familiar with some part of the world of nonlinguistic objects and also with linguistic sounds, especially phonemes. The primitive ability of the baby to achieve this familiarity, although not much more remarkable than the achievements of lower animals, is still very remarkable indeed, and more so, in the writer's opinion, than anything that comes later in his intellectual development. If this opinion is correct, then most of the struggle in constructing an ultraintelligent machine will be the construction of a machine with the intelligence of an ape.

The child later associates words with objects and activities, by implicit statistical inference: in fact the first words learned are surely

regarded by the child as properties of an object in much the same sense as the visual, olfactory, and tactual properties of the object. For example, if the child succeeds in pronouncing a word to an adequate approximation, and perhaps points in approximately the right direction or otherwise makes approximately the right gesture, then, statistically speaking, events are more likely to occur involving the object or activity in question; and, if the environment is not hostile, the events are likely to be pleasurable. Thus the words and gestures act as statistical index terms for the retrieval of objects, and the activation of processes. At a later stage of linguistic development, similar statistical associations are developed between linguistic elements themselves. The subject of statistical semantics would be concerned with all such statistical associations, between linguistic elements, between nonlinguistic and linguistic elements, and sometimes even between nonlinguistic elements alone.

A basic problem in statistical semantics would be the estimation of probabilities $P(W_i \mid O_j)$ and $P(O_j \mid W_i)$, where W_i represents a word (a clump of acoustic time series defined in a suitable abstract space, or, in printed texts, a sequence of letters of the alphabet with a space at both ends: admittedly not an entirely satisfactory definition), and O_j represents an object or an activity. $P(W_i \mid O_j)$ denotes the probability that a person, speaking a given language, will use the word W_i to designate the object O_j, and $P(O_j \mid W_i)$ is the probability that the object O_j is intended when the word W_i is used. Strictly, the estimation of probabilities is nearly always interval estimation, but, for the sake of simplicity, we here talk as if point estimation is to be used. The ranges of values of both i and j are great; the vocabulary of an educated man, in his native tongue, is of the order of 30,000 words and their simple derivatives; whereas the range of values of j is far far greater. The enormity of the class of objects is of course reducible by means of classification, which, in recognition, again involves a process of regeneration, just as does the recognition of a word.

An ideal statistical dictionary would, among other things, present the two probability matrices,

$$[(P(W_i \mid O_j)] \quad \text{and} \quad [P(O_j \mid W_i)]$$

(Compare Spärck Jones [*95*] and the discussion.) Such a dictionary would, apart from interdependences between three or more entities, give all the information that could be given, by a dictionary, for naming an object and for interpreting a word. Existing dictionaries sometimes indicate the values of the probabilities $P(W_i \mid O_j)$ to the extent of writing "rare"; and also the variations between subcultures are indicated ("archaic," "dialect," "slang," "vulgar," and so on). But let

us, somewhat unrealistically, imagine a statistical-dictionary maker who is concerned with a fixed subculture, so that the two probability transition matrices are fixed. One method he can use is to take linguistic texts, understand them, and thus build up a sample (f_{ij}), where f_{ij} is the frequency with which object O_j is designated by word W_i. Within the hypothetically infinite population from which the text is extracted, there would be definable probabilities $P(W_i)$ and $P(O_j)$ for the words and objects, and a *joint* probability $P(W_i \cdot O_j)$ crudely estimated by $f_{ij}/\Sigma_{ij}f_{ij}$. If these joint probabilities could be estimated, then the two probability matrices could be readily deduced.

We have now said enough to indicate the very close relationship that exists between statistical semantics, recall, and the retrieval of documents. In the remaining discussion in this section we shall restrict our attention to the question of retrieval of documents, including abstracts. This is a particular case of the retrieval of objects and the inauguration of processes, and the discussion brings out some of the theoretical difficulties of statistical semantics in a concrete manner.

A basic problem, bordering on semantics, is the estimation of the probability $P(D_j \mid W_i)$, where W_i represents a word, or index term, and D_j represents a document, or other object, and $P(D_j \mid W_i)$ denotes the probability that D_j represents a sought document, when W_i is an index term, and when it is not known what the other index terms are. Strictly speaking, the probability depends on the customer, but, for the sake of simplicity, it will be assumed here that the indexer of the documents, and all the customers, speak the same indexing language. The problem of estimating $P(D_j \mid W_i)$ is nontrivial to say the least [*48*], but let us imagine it solved for the present.

Next suppose that index terms $W_1, W_2, \ldots, W_m$ have been specified. Then we should like to be able to compute the probabilities $P(D_j \mid W_1 \cdot W_2 \cdot \ldots \cdot W_m)$, where the periods denote logical conjunction. One could imagine this probability to be estimated by means of a virtually infinite sample. Reasonable recall systems would be those for which (i) the probability that the document D_j will be recalled is equal to the above probability; (ii) the document D_j that maximizes the probability is selected; or (iii) the documents of highest (conditional) probability are listed in order, together with their probabilities. (Compare, for example [*35*], [*31*], p. 12, [*41*].)

In one of the notations of information theory [*26*, *67*],

$$\log P(D_j \mid W_1 \cdot W_2 \cdot \ldots \cdot W_m)$$
$$= \log P(D_j) + I(D_j : W_1 \cdot W_2 \cdot \ldots \cdot W_m) \qquad (5.1)$$

where $I(E : F)$ denotes the amount of information concerning E provided

by F, and is defined (for example [26, 16a]) as the logarithm of the "association factor"

$$I(E : F) = \log \frac{P(E \cdot F)}{P(E)P(F)} = \log \frac{P(E \mid F)}{P(E)} = \log \frac{P(F \mid E)}{P(F)} \qquad (5.2)$$

(The "association factor" as defined in refs [27], [31], and [41] is the factor by which the probability of one proposition is multiplied in the light of the other. It is used in a different sense, not as a population parameter, in Stiles [96].) The amount of information concerning E provided by F is a symmetrical function of E and F and is also called the "mutual information" between E and F, and is denoted by $I(E, F)$ when we wish to emphasize the symmetry. Notice that our "mutual information" is not an expected value as is, for example, the "relatedness" of McGill and Quastler [67]. Shannon [89] always used expected values.

If the index terms $W_1, W_2, \ldots, W_m$ provide statistically independent information concerning D_j (i.e., if $W_1, \ldots, W_m$ are statistically independent, and are also statistically independent given D_j), then

$$\log P(D_j \mid W_1 \cdot \ldots \cdot W_m) = \log P(D_j) + \sum_{r=1}^{m} I(D_j : W_r) \qquad (5.3)$$

The expected rate at which the individual index terms provide information concerning documents is

$$\sum_{i,j} P(W_i \cdot D_j) I(W_i, D_j)$$

conveniently denoted by $I(D : W)$ (compare [89], p. 90), but this does not allow for the expectation of the mutual information when several index terms are used. A knowledge of the expectations, for various values of m, would be relevant to the design of information-retrieval systems, since its antilogarithm would give some idea of the "cut-down factor" of an m-term request.

When one wishes to choose between only two documents, then the final log-*odds* are equal to the initial log-odds plus the sum of the "weights of evidence" or "log factors" (see [21] for the terminology here and cf. Minsky [73]).

It should be noted that Eq. (5.1) and (5.3) are just ways of writing Bayes' theorem, but this is not a stricture, since Bayes' theorem is likewise just a way of writing the product axiom of probability theory. It is suggestive to think of Bayes' theorem in a form that is expressible in one of the notations of information theory, since the various terms in Eqs. (5.1) and (5.3) might correspond to physical mechanisms, associa-

tive bonds between memory traces (see Section 6). The use of Eq. (5.3) might be described as the "crude" use of Bayes' theorem, or as a first approximation to the ideal procedure. It was used, for example, in [73], and I think also in [64]. It is a special case of discrimination by means of linear discriminants of the form

$$a_j + \sum_{r=1}^{m} b_{j,r} \qquad (5.4)$$

which have been used, for example, in the simplest class of perceptra [82], and in suggestions or experiments related to mechanical chess and chequers (draughts) (for example [33, 83, 18, 83a]).

One can write (5.4) in the form

$$a_j + \sum_{\mu} b_{j,\mu} \varepsilon_\mu \qquad (5.5)$$

where now the summation is over all words in the language, not just those that occur in the context, and ε_μ is defined as 1 if W_r does occur and as 0 otherwise. It is because of this mode of writing (5.4) that we call it a linear discriminant. It has been claimed [104] that the more general form, (5.4) or (5.5), is often much better than the crude use of Bayes' theorem, i.e., Eq. (5.3).

In order to estimate the second term on the right of Eq. (5.1), a very large sample would usually be required, and this is why it is necessary to make approximations. Successively better approximations can presumably be obtained by truncating the following series:

$$\log P(D_j \mid W_1, W_2, \ldots, W_m) = \log P(D_j) + \sum_r I(D_j, W_r) - \sum_{r,s} I(D_j, W_r, W_s) + \ldots \qquad (5.6)$$

$(r, s, t, \ldots = 1, 2, \ldots, m; r < s < t < \ldots)$, where the I's are "interactions of the first kind" as defined in the Appendix. If, for example, we were to truncate the series after the interactions of the second order (the J_2's), we would obtain a special case of the quadratic discriminant

$$a_j + \sum_{r=1,\ldots,m} b_{j,r} + \sum_{r,s=1,2,\ldots,m}^{r \neq s} c_{j,r,s} = a_j + \sum_{\mu} b_{j,\mu} \varepsilon_\mu + \sum_{\mu,\nu}^{\mu \neq \nu} c_{j,\mu,\nu} \varepsilon_\mu \varepsilon_\nu \qquad (5.7)$$

which, with optimal coefficients, would of course give a better approximation. (An example of the relevance of the quadratic terms, in analogous problems in learning machines, is in the evaluation of

material advantage in chess: the advantage of two bishops [33].) If we truncate Eq. (5.6) after the third-order interactions, we of course obtain a special case of a cubic discriminant, and so on.

An interesting class of problems arises when we ask: What are the optimal linear, quadratic, cubic, etc., discriminants, and how do we set about finding them? There is some discussion of this problem in [19] and in [85]. Here we merely make the obvious comment that, if the number of words is large, the number of coefficients increases rapidly with the degree, and optimization problems might be exceedingly difficult even for the cubic discriminant. Even without optimization, the work of estimating the interactions $I(D_j, W_r, W_s, W_t)$ would be enormous. It will be suggested, in Section 6, that the subassembly theory of the brain is capable of explaining, or at least of explaining away, how the brain can in effect embody these higher interactions as association bonds between sets of subassemblies. But in the present section we shall not consider biological processes.

Let us consider how, in principle, the various terms in Eq. (5.6) could be obtained. We should begin by taking a sample of *successful library applications*, each being of the form $(W_1, W_2, \ldots, W_m; D_j)$, meaning that the index terms $W_1, W_2, \ldots, W_m$ were used by the customer, and he was satisfied with the document D_j. If on a single occasion he was satisfied by more than one document, then, in this notation, that occasion would correspond to more than one successful library application. It should be remembered that we are assuming that all customers speak the same language. This assumption is in flagrant contradiction with the facts of life, but we assume it as an approximation in order to avoid complication. It should be noted that a sample of the kind mentioned here would form a useful part of any practical operational research on document retrieval.

We can now imagine the raw statistical data to be entered in a contingency table in $w + 1$ dimensions, where w is the number of index terms in use (the "size of the vocabulary"); w of the sides of the contingency table would be of length 2, whereas the remaining side would be of length d, the number of documents. It might be suggested that the way to enter the data in the table would be to regard each successful library application as a long vector

$$(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_w, j) \tag{5.8}$$

where ε_1 is 1 or 0 depending on whether the ith index term in a dictionary of index terms is one of the index terms, $W_1, \ldots, W_m$ that was used in the application; and so to put a tick in the cell (5.8) of the contingency table. This method of constructing the contingency table would be very

misleading, since there is a world of difference between missing out an index term and explicitly saying that the term is irrelevant to the sought document. This method of construction would be appropriate only if the entire vocabulary of index terms were presented to the customer to be used as a yes-no tick-off list. As R. A. Fairthorne has often pointed out, the idea of negation is not usually a natural one in indexing.

Instead, the above "successful library application" is better regarded as contributing to the "marginal total," denoted [47] by

$$n_{,,,\ldots,,,1,,,\ldots,,,1,,,\ldots,,,1,,,\ldots,,,j}$$

The meaning of this notation is this. Let

$$n_{\varepsilon_1\varepsilon_2\ldots\varepsilon_w j}$$

be the hypothetical entry in the contingency table in the cell given by (5.8); "hypothetical" since tick-off lists are not in fact used. In the above notation, each of the 1's, of which there are m, corresponds to the specification of an index term, and the acute accents indicate summations of $n_{\varepsilon_1\ldots\varepsilon_w j}$ over all the ε_i's that do not correspond to one of these m terms.

After a large amount of sampling, one would have good estimates for the values of many of the marginal totals of the "population contingency table," that is, the $(w + 1)$-dimensional array of population probabilities. The "principle of maximum entropy" [47, 56, 48] could then be used in principle for estimating all the $2^w d$ probabilities. The amount of calculation would be liable to be prohibitive, even if it were ultraparallel, although it might be practicable in analogous small problems such as the recognition of printed characters or phonemes.

It should be possible in principle to cut down the size of both the sample and the calculation by making use of the theory of clumps ("botryology") or of clusters. One of the benefits of such a theory would be that, by lumping together words into clumps, the dimensionality of the contingency table would be reduced.

The theory of clumps is still in its infancy (see for example [30, 35, 31, 41, 78, 75, 76]), and is necessarily as much an experimental science as a theory: this is why we prefer to call it "botryology." Research workers who use the term "cluster" rather than "clump" (see for example [77, 97, 81, 93, 78a][2]) seem to be concerned with the grouping of points that lie in a Euclidean space, and their methods tend to be fairly orthodox from the point of view of statistical methodology. In botryology the methods tend to be less orthodox, and it is sometimes

[2]The Editor mentions [78a], which I have not seen.

actually desirable that the clumps should overlap, both in applications to information retrieval and in potential application to neural nets. Nevertheless the two theories might be expected eventually to merge together.

Let us consider how botryology might be applied for finding clumps of associated index terms and "conjugate" clumps of documents. (The method could also be applied to the categorization of diseases, by replacing the index terms by symptoms and the documents by people.) Let there be w index terms, and d documents. Let f_{ij} be the frequency with which index term i occurs in document j, and consider the w by d matrix

$$F = (f_{ij})$$

Various botryological computations with F have been suggested in the references: the one proposed here is closest to that of Needham [*76*], who, however, was concerned with a square symmetric matrix of frequencies of co-occurrence of index terms, and who did not use logarithms or "balancing," as described below.

First replace the matrix F by the matrix $[\log (f_{ij} + k)]$, where k is a small constant (less than unity). A reason for using the logarithm is that we are proposing to use additive methods and a sum of log-frequencies is a log-likelihood. The addition of the small constant k to the frequencies is necessary to prevent zeros from going to minus infinity, and can be roughly justified for other reasons (see for example [*58*], [*25*], p. 241, or [*48*]). This modified matrix is now "balanced" in the following sense.

By balancing an arbitrary matrix we mean adding $a_i + b_j$ to cell (i, j) $(i, j = 1, 2, \ldots)$ in such a manner that each row and each column adds up to zero. It is easy to show that the balanced matrix is unique, and that the balancing constants can be found by first selecting the a_i's to make the rows add up to zero, and then selecting the b_j's to make the columns add up to zero. The column balancing does not upset the row balancing. For a symmetric matrix the row-balancing constants are equal to the column-balancing constants. In what follows, instead of balancing the matrix it might be adequate to subtract the mean of all the entries from each of them.

Let B be the result of balancing the matrix $[\log(f_{ij} + k)]$. Consider the bilinear form $b = x'By$, where x is a column vector consisting of $+1$'s and -1's, and the prime denotes transposition. We now look for local maxima of b in the obvious manner of first fixing x, perhaps randomly, and finding y to maximize b (i.e., taking $y = \text{sgn } B'x$), and then fixing y and finding x to maximize b (i.e., taking $x = \text{sgn } By$), and so on iteratively. The process terminates when the bilinear form takes the same value twice running. The process would lead to the separation of

the words into two classes or large clumps, and two conjugate clumps of documents.

Consider one of the two smaller matrices obtained by extracting the rows and columns from B, corresponding to a clump and its conjugate. Balance this smaller matrix, and find a local maximum of its bilinear form. This procedure will split our clump into two smaller clumps, and will simultaneously split the conjugate clump. In this manner we can continue to dichotomize our clumps until they are of approximately any desired size. The whole collection of clumps would form a tree.

Actually, it is desirable that clumps should overlap in some applications to information retrieval, and this can be achieved by means of a slight modification of the above procedure, in which the "large" clumps are made larger still. That is, in place of taking all the $+1$'s in w as a clump, one could take all the components in $B'x$ algebraically greater than some negative threshold; and, in the conjugate clump, all the components in By above some negative threshold.

The effect of this botryological procedure is to induce a partially ordered structure each of whose elements is a clump of index terms together with its conjugate clump of documents.

Having obtained the partially ordered set of clumps, one could apply the methods described in [*48*], which, however, have not been completely worked out, in order to make estimates of $I(i, j)$ when f_{ij} is too small for the estimate $\log f_{ij} - \log f_{i\prime} - \log f_{\prime j}$ to be usable (for example when $f_{ij} = 0$ or 1). (We have written $f_{i\prime}$ and $f_{\prime j}$ for the total frequencies of W_i and D_j.) Hopefully, the higher-order mutual information (interaction) $I(W_1, W_2, \ldots, W_m \mid D_j)$ could be estimated in a similar manner.

Another conceivable method for associating documents with index terms would be in terms of the eigenvectors of BB' and of $B'B$, where the primes still indicate transposition. By a theorem of Sylvester, the eigenvalues of $B'B$ are the same as those of BB', together with $d - w$ zeroes, if $d \geq w$. We can use the nonzero eigenvalues in order to pair off the two sets of eigenvectors, and we could order each of the two sets of eigenvectors in the order of the magnitudes of the eigenvalues. Then we could associate with the ith index term the ith component of the normalized eigenvectors of BB', and with the jth document the jth component of the corresponding w eigenvectors of $B'B$. This would associate a w-dimensional vector with each index term and with each document. The relevance of index term i to document j could now be defined as the correlation coefficient between the two associated vectors. An approximate relationship between relevance and mutual information could then be found experimentally, and we could then apply Eq. (5.1) for document retrieval. The amount of calculation required for the

application of this method would be exceedingly great, whereas the clumping algorithm mentioned before could perhaps be carried out on a computer of the next generation.

6. Cell Assemblies and Subassemblies

Suppose that one wishes to simulate psychological association and recall on a machine. We restrict our attention to *the recall of one word when m other words are presented*, but most of the discussion can be adapted, in an obvious manner, to the recall of a concept given various attributes, or to the retrieval of a document, given various index terms. The discussion could be modified in order to cover the case when the words are presented serially and form a Markov chain, this being a well-known approximate model for the prediction of words in a language text (cf. [*88*]). For the sake of simplicity, we shall ignore problems of syntax, so that our discussion will be in this respect more pertinent to methods of information retrieval based only on index terms than to the full problem of recall. This limited problem is difficult enough for the present, and is I think a necessary preliminary to any more ambitious discussion of recall in general.

If there are w words in the vocabulary, there are potentially $w(w - 1)/2$ associations of various strengths between pairs of words. (*Kinds* of association are here being ignored.) The process of recall, in this example, is that of selecting the word, A, that is in some sense most associated with the m words $A_1, A_2, \ldots, A_m$ which have been recently inserted at the input of the computer. In the usual problem of information retrieval A would be a document and $A_1, A_2, \ldots, A_m$ would be index terms, and the discussion of the previous section is all relevant to the present problem.

The difficulty of making the probability estimates [*48*] provides some of the explanation of why men are not entirely rational in their probability estimates and in their recall. It is possible, for example, that, for men, the probability of retrieval of a word is approximated by only a few terms of Eq. (5.6) of the previous section. An ultraintelligent machine might be able to use more terms of the equation, since it might be able to speed up the calculations by invoking the electronic computer with which it would be in close relationship (cf. [*33*]).

Russell and Uttley [*102*] suggested that a time delay might be the neural embodiment of the amount of information in a proposition, $I(H) = -\log P(H)$, and that this would make conditional probabilities easily embodiable, since the difference between two time delays is itself a time delay. As point out in [*38*], this idea extends at once to mutual information, log-odds, weights of evidence, and tendency to

cause. But of course time delay is only one of many possible physical representations of a real variable, and others could be suggested in terms of synaptic facilitation. In view of the complexity of the brain, it is quite probable that more than one representation is used, and this would give greater scope for adaptability. One must not be overready to apply Ockham's lobotomy. As in other complex systems, many theories can contain elements of the truth. Economists are familiar with this principle.

We return now to our problem of information retrieval.

Suppose that $w = 30{,}000$ and that some acceptable method were found for estimating the mutual information between each pair of the 30,000 words. Then it will still be hardly practicable to list the 450 million answers in immediately accessible form in a machine that is not ultraparallel. Instead it would be necessary to put the words that have appreciable association with a given word, A, into a list of memory locations, called say the A list. Each word in each list must have the strength of the association (the logarithm of the association factor) tagged to it. Many of the lists would be very long. The process of recall involves the collation of the words in the lists corresponding to recent input words, together with some further arithmetic. Collation is a slow process, and it is tempting to ask whether it would be more economical to simulate the process of recall by means of an artificial neural network, or at any rate by means of ultraparallelism. The use of artificial associative memories is a step in this direction, but so far only a small one (for example [*60*, *65*]). For purposes of information retrieval, which in effect is what we are discussing, it might be worth while to design computers that are not ultraparallel but have extremely rapid collation as a special feature. Such computers would be very useful for information retrieval by means of index terms, but when the words are strongly interdependent statistically, as in ordinary language, a machine using artificial neural nets seems intuitively to hold out more promise of flexibility. (See also the discussions of "higher-order interactions" later in this section).

If each word were represented by an artificial neuron, or otherwise highly localized, it would take too long to set up the associations, unless there were $w(w - 1)$ association fibers built in, and this would be very expensive in equipment. Moreover, it is not easy to see how more than a small fraction of such a machine could be in operation at any one time, so that there would be a great wastage of potential computation power. For these reasons, a machine with "distributed memory" seems more promising. As Eccles says ([*16*], p. 266), "Lashley argues convincingly that millions of neurons are involved in any memory recall, that any memory trace or engram has multiple representation; that each neuron

or even each synaptic joint is built into many engrams" [*61*]. A further relevant quotation, from [*34*], is:

"An interesting analogy is with the method of superimposed coding, of which Zatocoding is an example. This is a method of coding of information for information-retrieval purposes. Suppose we wish to identify a document by means of m index terms. Each term is represented by means of ν punched holes in a card containing N locations each of which can be punched or not punched. [For each of the index terms] we may select ν locations out of the N at random [to punch]. The representation of the joint occurrence of m index terms is then simply the Boolean sum of the m individual punchings of ν locations each. . . . In the application to information retrieval if we extract all the cards punched in the ν locations corresponding to any given term, we may get some cards that are irrelevant by chance. If N is large, and ν is suitably selected, mistakes need seldom occur. In fact it is natural to arrange that

$$(1 - \nu/N)^m \approx \tfrac{1}{2}$$

i.e.,

$$\nu \approx (1 - 2^{-1/m})N$$

This must be the best value of ν since to have half the holes punched gives the largest variety of possible punchings.

"By analogy, Nature's most economical usage of the brain would be for a reasonable proportion of it to be in operation at any one time, rather than having one concept, one neuron." Instead, each neuron would occur in a great many distinct circuits, and would not be indispensable for any of them.

Such an analogy can at best give only a very rough idea of what goes on in the brain, which is an ultradynamic system as contrasted with a collection of punched cards. (The analogy would seem a little better if, instead of taking the Boolean sum, a threshold were used at each location.) But if we take $m = 20$, on the grounds that the game of "twenty questions" is a reasonably fair game, we find that the representation of a word occupies say a thirtieth of the neurons in the cortex. It must be emphasized that this is not much better than a guess, partly because it is based on a very crude optimality principle. But it is not contradicted by the experiments of Penfield and others (for example [*80*], p. 117) who found that the electrical stimulation of a small area on the surface of the cortex could inhibit the recall of a fraction of the subject's vocabulary. (For further references, see Zangwill [*108*].) For it is entirely possible that a large subnetwork of neurons could be inhibited, and perhaps even sparked off, by stimulation at special points.

Among the theories of distributed memory, the "cell assembly" theory is prominent, and, as stated in the previous section, a modified

form of this theory will be adopted here. The meaning and point of the theory can be explained in terms of its applications to the linguistic activities of the brain, although the theory is usually discussed in a more general context. There are some advantages in discussing a special case, and some generalizations will be obvious enough. A cell assembly is assumed to consist of a great number of neurons, which can all be active at least once within the same interval of about a quarter to half a second. For simplicity we shall generally take the half-second estimate for granted. An assembly reverberates approximately as a unit, and, while reverberating, it tends to inhibit the remainder of the cortex, not neuron by neuron, but enough so that no other assembly can be very active during the same time interval. A word, or a familiar phrase, is often represented by an assembly, and, more generally, an assembly usually corresponds in Hebb's words, to a "single element of consciousness." But the consciousness might habituate to assemblies that occur very frequently.

It will be assumed in this paper that there are also subassemblies that can be active *without* dominating the whole cortex, and also that when an assembly becomes fatigued and breaks up it leaves several of its own subassemblies active for various lengths of time, from a second to several minutes, and typically about ten seconds. Each subassembly would consist of a smaller group of neurons than an assembly, but with greater relative interconnectivity. The subassemblies might in their turn break up into still smaller groups of still greater relative interconnectivity and of greater "half-lives." These could be called subsubassemblies, etc., but we shall usually use the term "subassembly" generically to include subsubassemblies, etc. When an assembly gains dominance for a moment it is approximately completely active, when the subject is wide awake. The process is assumed to be one of approximate regeneration. It is not exact regeneration for if it were there would be no learning. Probabilistic regeneration might often be represented by the *degree* of activity of an assembly. This degree of activity will be carried forward by the subassemblies, so that the benefits of probabilistic regeneration, as described in a previous section, will be available. Also the activity is less, and the assembly is somewhat smaller, when the subject is sleepy or dreaming, but the activity is assumed to be nearly always enough for the assembly to have a definite identity, except perhaps in dreamless sleep. When the subject is nearly asleep, there might be frequent intervals of time when there is no active assembly.

The association between two assemblies could be largely embodied in the subassemblies that they have in common.

When a man is in a sleepy condition, an assembly need not be followed

by another consciousness-provoking assembly for a short time. In that case, the assembly A might recover from fatigue and be reactivated by the subassemblies that it itself had left in its wake when it last fired. This would account for the occasional repetitivity of thought when one is sleepy. The hypothesis is not that the assembly reverberates for longer than usual, but that it is liable to reactivate because there has not been enough activity to disperse its subassemblies. The subassemblies themselves, both in sleepiness and in dreams, have lower activity than in wakefulness, so that, when one wakes up, the memory and atmosphere of dreams would be easily erased. When dreaming there is perhaps not enough energy in the cortex to sustain many full assemblies so that the subassemblies would be less inhibited than in wakefulness. It might well be that there are far more subassemblies active during sleep, and they would form arrangements having less logical cohesion and higher entropy. This would explain the remarkable rate at which visual information can be internally generated during dreams; and the incomplete regeneration of full assemblies would explain the *non sequitur* and imaginative nature of dreams.

In the same spirit, if assemblies correspond to conscious thoughts, *it might well be that subassemblies correspond to unconscious and especially to preconscious thoughts, in the wakeful state as well as in sleep.*

What gives the assemblies their semipermanent static structures, corresponding to long-term memory, is assumed, following Hebb, to be the pattern of strengths of synaptic joints throughout the cortex. The physical counterpart of learning is the variation of these strengths. We have already conjectured that the number of possible states of any synaptic joint is small enough to justify calling the strength a "discrete variable." This assumption makes it easier to understand how memories can be retained for long periods, and how the identities of assemblies can be preserved.

We assume that the strength of a synapse, when not in use, occasionally *mutates* in the direction of some standard value. This mechanism would explain the gradual erosion of memories that have not been recalled, and would also help to prevent all synapses from reaching equal maximal strength, which would of course be disastrous. Equally, the increase in strength of a synapse when its activation leads to the firing of a neuron can reasonably be assumed to be a mutation and only probabilistic. The number of synapses is so large that it might well be sufficient for only a small fraction of them to mutate when they contribute to the firing of a neuron. *This hypothesis would also help to explain why all synapses do not reach maximum strength.*

Even when an assembly sequence is frequently recalled, some of the strengths of the relevant synapses would nevertheless have mutated

downwards, so that some of the many weaker subassemblies involved in the assembly sequence would have become detached from the structure of the assembly sequence. Thus the structure of a frequently used assembly sequence, used only for recall and not for building into fantasy or fiction, would tend to become simplified. In other words, detail would be lost even though what is left might be deeply etched. Thus the corresponding memory would tend to become stereotyped, even in respect of embellishments made to it after the first recording.

It is interesting to consider what enables us to judge the time elapsed since a memory was first inscribed. Elapsed time seems introspectively to be recorded with roughly logarithmic accuracy: the least discernible difference of a backward time estimate is perhaps roughly proportional to the time elapsed, not allowing for the "cogency" of the recall, that is, not allowing for the interconnections and cross-checks in the recall. This conjecture, which is analogous to the Weber-Fechner law, could be tested experimentally. An aging memory suffers from a gradual loss of "unimportant" detail. If, on the other hand, we recall an item repeatedly, we preserve more of the detail than otherwise, but we also overlay the memory with additional associations to assemblies high up in the hierarchy. *We can distinguish between "reality" and imagination* because a memory of a real event is strongly connected to the immediate *low-order* sensory and motor assemblies. As a memory ages it begins to resemble imagination more and more, and the memories of our childhood are liable to resemble those of a work of fiction.

One of the advantages that an ultraintelligent machine would have over most men, with the possible exception of millionaires, would be that it could record all its experiences in detail, on photographic film or otherwise, together with an accurate time-track. This film would then be available *in addition* to any brain-like recordings. Perfect recall would be possible without hypnotism!

As pointed out by Rosenblatt ([*82*], p. 55), a permanent lowering of neural firing thresholds would be liable to lead to all thresholds becoming minimal, unless there were a "recovery mechanism." He therefore prefers the more popular theory of synaptic facilitation, which we are using here [*15, 54*]. Although there are far more synapses than neurons, a similar objection can be raised against this theory, namely, too many synapses might reach maximal facilitation, especially if we assume a cell assembly theory. This is why we have assumed a mutation theory for synaptic strengths. In fact, we assume both that a synapse, when not in use, mutates downwards, with some probability, and also, that when it has just been used, it mutates upwards, with some probability. The higher the strength at any time, the greater the probability of mutating downwards when not used, and the smaller the probability of mutating

upwards when used. It is neither necessary nor desirable that every synapse should increase its strength whenever it is used. The enormous number of neurons in an assembly make it unnecessary, and the frequent uses of the synapses make it undesirable. After a certain number of uses, an assembly does not need any further strengthening.

A sentence lasting ten seconds would correspond to an *assembly sequence* of about twenty assemblies. Hebb ([*54*], p. 143) says that the apparent duration of a "conceptual process" in man is from one to five or ten seconds. The expression "conceptual process" is of course vague, and the discussion is here made somewhat more concrete by framing it in terms of linguistic activity. A phoneme, when it is part of a word, perhaps corresponds to a subassembly, and there will be many other subassemblies corresponding to other properties of the word, but only a fraction of these will remain active when the assembly breaks up.

Which assembly becomes active at the next moment must depend on the current sensory input, the current dominant assembly, and the currently active subassemblies. *Indirectly, therefore, it depends on the recent assembly sequence*, wherein the most recent assemblies will have the greatest influence. It also depends of course on the semipermanent static storage, the "past history." Well-formed assemblies will tend to be activated by a small fraction of their subassemblies; this is why it is possible to read fast with practice: it is not necessary to observe all the print. Memory abbreviates.

An example that shows how the activation of an assembly can depend on the previous assembly sequence is the recall of a long sequence of digits, such as those of π. A. C. Aitken and Tom Lehrer, for example, can repeat several hundred digits of π correctly. If we assume that there is one assembly for each of the ten digits 0, 1, . . ., 9, then it is clear that the next assembly to be activated must depend on more than just the previously active assembly. If there is no hexanome (sequence of six digits) that is repeated in the first 500 digits of π, then *one* method of remembering the 500 digits in order is to memorize a function of hexanomes to mononomes. Then any six consecutive digits would uniquely determine the next digit in this piece of π; for example, the digit 5 is determined by the hexanome 415926.

Let us consider how the subassembly theory would account for this. For the sake of argument, we shall ignore the strong possibility that a calculating prodigy has an assembly for say each of the hundred distinct dinomes, and continue to assume one assembly for each of the ten digits. (*The argument could be modified to allow for other possibilities.*) *We take it for granted that the subject (Aitken) is in the psychological* "set" corresponding to the recitation of the digits of π. Suppose that the assembly corresponding to the digit i has subassemblies $s(i, 1)$, $s(i, 2)$,

. . ., and that these symbols correspond to subassemblies of successively shorter "half-lives." Then, provided that the digits are recited by Aitken at a constant rate, one set of subassemblies that would activate the assembly corresponding to 5 would be of the form $s(4, 1)$, $s(4, 2)$, . . ., $s(4, n_{4,6})$; . . . ; $s(6, 1)$, $s(6, 2)$, . . . , $s(6, n_{6,1})$, where $s(i, n_{i,j})$ is the next subassembly (belonging to assembly i) to become extinguished after j "moments of time." If at least one subassembly of each assembly is extinguished at each moment within the first six moments after the assembly is extinguished, then this theory could account for the possibility of the recitation. For, at any given moment, the active subassemblies would uniquely determine the next assembly to be activated. If the recitation were slowed down by a moderate factor, then there would still be enough clues for the unique determination of the successive digits. In fact a knowledge of the maximum slow-down factor would give quantitative information concerning the numbers and durations of activation of the subassemblies.

There is an analogy between cell assemblies and the gel that can form in a polymerization reaction. (See Flory [*17*] for a comprehensive discussion of polymerization, or [*45*] for a short self-contained description of some mathematical theory that might also be relevant to cell assemblies.) The gel is often regarded as a molecule of infinite size, but there can be other largish molecules present simultaneously, analogous to the subassemblies. Polymerization is not as dynamic as cerebral activity, so the analogy is imperfect, but it is instructive since it shows the plausibility of subassemblies.

A theory that does some of the work of the subassembly theory is the theory of "primed neurons" ([*43*], p. 506 and [*71*]). We quote (from the former reference): "After an assembly has just been extinguished, many of its neurons will have received subthreshold activation without having fired. Milner calls them 'primed neurons'. . . . A primed neuron may be regarded as the opposite of a refractory one. Therefore, in virtue of 'temporal summation' for neurons, parts of a recently extinguished assembly will be primed, so that it will be easily reactivated during the next few seconds. This is an explanation of short-term memory different from that of reverberatory circuits; but an activated assembly must itself reverberate. Milner assumes that the effect of priming dies away after a few seconds. But I think it would be useful to assume that the time constant can vary greatly from neuron to neuron since this may help to explain our sense of duration, and also medium-term memory. Here, as elsewhere, other explanations are possible, such as the gradual extinction of small reverberating circuits within assembles." (The last remark is a reference to subassemblies; see also [*41*].)

The subassembly theory seems to be a more natural tool than that of primed neurons, for the purpose of explaining the sequence of firing of assemblies although both might be features of the brain. One would expect subassemblies to exist, since the density of connectivity in an assembly would be expected to vary from place to place in the cortex. Subclumps of high connectivity in a network would be expected to reverberate longer than those of low connectivity. Although it could be argued that highly connected subclumps should become exhausted more quickly, it should be observed that the synapses in these subclumps will tend to be stronger than where the connectivity is low. It is therefore natural to assume that the subclumps correspond to subassemblies.

It might turn out that the theory of primed neurons will be sufficient to explain the workings of the brain, without the assumption of subassemblies, but the latter theory gives the kind of discrete representation that fits in well with the notion of probabilistic regeneration.

The theory of subassemblies is so natural for any large partly random-looking communication network (such as that of a human society) that it tempts one to believe, with Ashby ([*4*], p. 229), that a very wide class of machines might exhibit intelligent behavior, provided that they have enough interconnectivity and dynamic states. Machines certainly need *some* design, but it is reasonable to suppose that money and complication can be traded for ingenuity in design. For example, a well-designed machine of say 10^9 components might be educable to ultra-intelligence, but a much more carelessly designed machine of say 10^{13} components might be equally good.

That some design is necessary can be seen from one of the objections to the cell assembly theory as originally propounded by Hebb. Hebb did not originally assume that it was necessary to assume inhibition, and Milner pointed out that, without inhibition, the assemblies would fill the whole cortex. Ultimately there could be only one assembly. Either inhibition must be assumed to exist, as well as excitation, or else the assemblies would have to be microscopically small in comparison with the cortex. The latter assumption would be inconsistent with "distributed memory." Milner accordingly assumed that neurons tend to inhibit those near them. Therefore one may picture an assembly as a kind of three-dimensional fishing net, where the holes correspond to inhibited neurons.

The simplest model would assume that each fishing net (assembly) spans the entire cortex, or perhaps only the entire association cortex, or perhaps also other parts of the brain [57]. *In future, mainly for verbal simplicity, we use the word "cortex" unqualified.* There is a need for some mathematical theorems to show that a very large number of

distinct assemblies could exist under reasonable assumptions for the parameters that describe connectivity. It is reasonable to conjecture that the thinness of the cortex is a relevant parameter, or rather the "topology" that is encouraged by the thinness. The dimensions of the cortex, if straightened out, would be about 50 cm by 50 cm by 2 mm ([*90*], pp. 32 and 34). It is possible that the assembly theory would become impossible if the cortex were much "thicker." If we cannot treat the problem mathematically, perhaps we should experiment with an artificial neural net of neural dimensions approximately $50 \times 10{,}000 \times 10{,}000$, but smaller-scale experiments would naturally be tried first. There must surely be some advantage in having thin cortices, otherwise people would have thicker ones. It seems unlikely that the brain contains many useless residuals of evolutionary history. Hence the anatomy of the brain is very relevant to the design of the first ultraintelligent machine, but the designer has to guess which features have important operational functions, and which have merely biochemical functions.

Since it is not known what values of the parameters are required for the intelligent operation of a neural net, it is possible only to guess which features of the cortex are most relevant for the design of an ultra-intelligent machine. The feature of a good short-term memory ("attention span"), of the order of 20τ, where τ is the active time of a single assembly, is certainly essential for intelligence. (In a machine τ need not be approximately half a second.) It might even be possible to improve on the performance of a brain by making the average duration of the sequence somewhat greater than 20τ. But there must be a limit to the useful average duration, for a given cost in equipment. This limit might be determined by the fact that the longer an assembly sequence the smaller must be the average size of the assemblies; but is more likely to be determined by the fact that the complexity of concepts can be roughly measured by the durations of the assembly sequences, and beyond a certain level of complexity the brain would not be large enough to handle the *relationships* between the concepts. (In a more precise discussion the duration would be interpreted as a kind of "half-life.")

When guessing what biological features are most relevant to the construction of an ultraintelligent machine, it is necessary to allow for the body as a whole, and not just the brain: an ultraintelligent machine would need also an input (sensorium) and an output (motorium). Since much of the education of the first ultraintelligent machine would be performed by a human being, it would be advisable for the input and output to be intuitively tangible. For example, the input might contain a visual and a tactual field and the output might control artificial limbs. In short the machine could be something of a robot. The sensorium and motorium might be connected topographically to parts of

the two surfaces of the disk that represents the cortex. Many other decisions would have to be made concerning the design, even before any really useful experiments could be performed. These decisions would concern qualitative details of structure and also the values of quantitative parameters. The need for further theory is great, since, without advances in theory, the amount of experimentation might be prohibitive. Even if the values of the parameters in the cerebral cortex were known [*90*], theory would be required in order to decide how to scale them to a model with fewer components. A very tentative example of some quantitative theory is given near the end of the present section.

It has been argued [79] that the cortex seems to be under the control of a more centrally placed subcortical region, partly in the diencephalon, "not in the new brain but in the old" ([80], p. 21).[3] *Penfield calls the partly hypothetical controlling region the "centrencephalic system." It* seems that consciousness is likely to be associated with this system. A natural inference of the hypothesis that consciousness is associated with the old brain is that the lower animals have consciousness, and can experience "real metaphysical pain," an inference natural to common sense but disliked by some experimentalists for obvious reasons: they therefore might call it meaningless.

Sometimes Penfield's theory is considered to be inconsistent with Hebb's, but in the present writer's opinion, *the assembly theory is made easier to accept by combining it with this hypothesis of a central control.* For the following mechanism suggests itself. The greater the amount of activity in the cortex, the greater the number of inhibitory pulses sent to all currently inactive parts of the cortex by the centrencephalic system. This negative feedback mechanism would prevent an assembly from firing the whole cortex, and would also tend to make all assemblies of the same order of size, for a given state of wakefulness of the centrencephalic system. This in its turn would be largely determined by the condition of the human body as a whole.

This "assembly theory, MARK III," as we may call it (taking a leaf out of Milner [*71*]), has two merits. First, it would allow a vastly greater class of patterns of activity to assemblies: they would not all have to have the pattern of a three-dimensional fishing net, filling the cortex. This makes it much easier to accept the possibility that a vast variety of assemblies can exist in one brain, as is of course necessary if the assembly theory is to be acceptable. A second, and lesser, merit of the modified theory is that a single mechanism can explain both the control of the "cerebral atomic reactor" and degrees of wakefulness, and perhaps of psychological "set" also. Finally, the theory will shortly be seen

[3]Zangwill gives earlier references in his interesting survey [*108*].

to fit in well with a semiquantitative theory of causal interactions between assemblies.

It is proposed therefore that our artificial neural net should be umbrella-shaped, with the spikes filling a cone.

During wakefulness, most assemblies will have a very complicated structure, but, during dreamless sleep, the centrencephalic system will become almost exclusively responsible, directly and indirectly, for the activity in the cortex, taking for granted of course the long-term or "static" structure of the cortex. The input from the cortex to the centrencephalic system will, as it were, be "reflected back" to the cortex.

The assumption is that the excitation put out by the centrencephalic system has the function of encouraging cortical activity when it is low, and discouraging it when it is high. Under a wide class of more detailed models, the amount of activity will then have approximately simple harmonic amplitude when other input into the cortex is negligible. Since we are assuming that the duration of a cell assembly is about half a second, following Hebb, it is to be expected that the period of this simple harmonic motion will also be about half a second. *This would explain the delta rhythm* ([*103*], p. 167) which occurs during sleep. Apparently, very rhythmic assemblies do not correspond to conscious thought. To some extent this applies to all assemblies that are very frequently used. Consciousness is probably at its height when assemblies *grow*.

In order to explain the alpha rhythm, of about five cycles per second, when the eyes are closed and the visual imagination is inactive, along similar lines, we could assume that "visual assemblies" have a duration of only about a fifth of a second. This would be understandable on the assumption that they are on the whole restricted to the visual cortex, i.e., to a smaller region than most other assemblies (cf. Adrian and Matthews [*2*]).

We have assumed that, when no assembly is active, the centrencephalic system encourages cortical activity, so that, at such times, the various current active subassemblies will become more active. This process will continue until the activity reaches a critical level, at which moment the neurons not already active are on the whole inhibited by those that are active, including those in the centrencephalic system. This is the moment at which, by definition, an assembly has begun to fire. If this happens to be a new assembly, then the interfacilitation between its subassemblies will establish it as an assembly belonging to the repertoire of the cortex. This will happen whenever we learn something new or when we create a new concept.

The newborn child has certain built-in tendencies, such as the exercise of its vocal organs. We assume that there are pleasure centers in the

brain, whose function is reinforcement, and that they are usually activated when there is a "match" between a sound recently heard and one generated by the vocal organs. The matching could be done by a correlation mechanism, which in any case is apparently required in order to recognize the direction of a sound. E. C. Cherry [13] points out the need for this, and also the possibility of its more general application (see also [57, 63]). Also the child is rewarded by attention from its parents when it pronounces new phonemes for the first time. Thus one would expect assemblies to form, corresponding to the simplest correctly pronounced phonemes. The phonemes in agricultural communities might be expected to be influenced by the farm animals. Assemblies corresponding to syllables and short words would form next, apart from the words that were negatively reinforced. Each assembly representing a word would share subassemblies with the assemblies that represent its phonemes. An assembly for a word would also have subassemblies shared with nonlinguistic assemblies, such as those representing the taste of milk, and, more generally, representing experiences of the senses, especially at the nine apertures, where the density of neurons is high for evolutionary reasons. And so, gradually, the largely hierarchical structure of assemblies would be formed, the lowest levels being mostly closely connected with the motorium and also with the sensorium, especially where the surface neural density is high.

It is interesting to speculate concerning the nature of the associations between cell assemblies. We shall suppose that there is some measure of the strength of the association from one cell assembly, F, to another one, A, or from an assembly sequence F to the assembly A. Assuming the subassembly theory, this association will be largely embodied in the strength of the association to A from the subassemblies left behind by F, and will depend on the degrees of activation of the subassemblies and on the current psychological "set." A few distinct but related formulas suggest themselves, and will now be considered. In these formulas we shall take for granted the degrees of activation and the psychological set, and shall omit them from the notation.

The first suggestion is that the strength of the association from F to A should be measured by $I(A : F)$, as in the discussion of information retrieval in Section 5. If F is the assembly sequence $A_1, A_2, \ldots, A_m$, and if these assemblies supply statistically independent information, we have, by Eq. (5.3):

$$\log P(A \mid A_1 \cdot A_2 \cdot \ldots \cdot A_m) = \log P(A) + \sum_{r=1}^{m} I(A : A_r)$$

It could then be suggested that the term $\log P(A)$ is represented by the

strength of the connectivity from the centrencephalic system to A. Actually it is unlikely that the assemblies will supply statistically independent information, and it will be necessary to assume that there are interaction terms as in Eq. (5.6). We would then have an explanation of why the next assembly that fires, following an assembly sequence, is often the one that ought to have the largest probability of firing in a rational man. More precisely, the terms $I(A : A_r)$ corresponding to the most recently active assemblies will be represented with larger weights. Consequently, when we wish to recall a memory, it pays to hold in mind all the best clues without the intervention of less powerful clues.

An objection to the above suggestion is that it is necessary to add a constant to $\log P(A)$ to make it positive, and then the neurophysiological "calculation" of the strength of the association from the centrencephalic system would be ill-conditioned. Accordingly we now consider another suggestion.

One of the distinctions between the action of the brain and document-retrieval systems is that the brain action is considerably more dynamic. The activity of the assemblies constitutes an exceedingly complicated causal network. It is natural to consider whether the causal calculus [39] might be applicable to it.

Reference [39] contains two immediately relevant formulas, namely,

$$Q(E : F) = \log \frac{1 - P(E \mid \bar{F})}{1 - P(E \mid F)}$$

the tendency of F to cause E ($\bar{F}$ denotes "not F"), also described as "the weight of evidence against F if E does not occur"; and

$$K(E : F) = \log \frac{1 - P(E)}{1 - P(E \mid F)} = -I(\bar{E} : F)$$

the "intrinsic" tendency of F to cause E. In both formulas, the laws of nature, and the state of the world immediately before the occurrence of F, are taken for granted and omitted from the notation. Like the mutual information, both Q and K have the additive property

$$Q(E : F \cdot G) = Q(E : F) + Q(E : G \mid F)$$

$$K(E : F \cdot G) = K(E : F) + K(E : G \mid F)$$

Moreover

$$Q(E : F \cdot G) = Q(E : F) + Q(E : G)$$

$$K(E : F \cdot G) = K(E : F) + K(E : G)$$

when F and G are "independent causes" of E. This means that F and G are statistically independent, and are also statistically independent

given *not* E. This definition of independent causes, extracted from [*39*], was seen to be a natural one by the consideration of a firing squad: E is the event that the victim is shot, F and G are the events of shooting by two marksmen; and part of the given information, taken for granted and omitted from the notation, is that the sergeant at arms gives the order to fire.

We now take F as the firing of an assembly or assembly sequence, also denoted by F, and we take E as the firing of the assembly A. The suggestion is that Q or K is a reasonable measure of the strength of the association from F to A. We then have additivity in so far as the components of F, assemblies or subassemblies, have independent tendencies to cause A to fire. Otherwise various interaction terms can be added, and can be expressed in various ways, for example,

$$K(E : F \cdot G) = K(E : F) + K(E : G) + I(F : G) - I(F : G \mid \bar{E})$$

The "causal force," $K(E : F)$, tends to activate A, but the assembly that is activated will not be the one that maximizes $K(E : F)$, but rather the one that maximizes $P(E \mid F)$. This can be achieved by assuming that the centrencephalic system applies a "force" $-\log[1 - P(E)]$. [This will always be well approximated simply by $P(E)$.] The resultant force will be $-\log[1 - P(E \mid F)]$ and increases with $P(E \mid F)$ as it should. We see that $K(E : F)$ appears to be more logical than $Q(E : F)$ for our purpose, since it would be more difficult to see how the centrencephalic system could apply a "force" equal to $-\log [1 - P(E \mid \bar{F})]$ to A.

If there exists no E for which $-\log[1 - P(E \mid F)]$ exceeds some threshold, then a *new* assembly will be activated, or else the next thought that occurs will be very much of a *non sequitur*.

It could be asked, what is the advantage of using $K(E : F)$ rather than $-\log[1 - P(E \mid F)]$, as a measure of the strength of the association from F to A? (In the latter case the centrencephalic system would not need to make a contribution.) Two answers can be given: first that, if $P(E \mid F) = P(E)$, then F should have *no* tendency to cause A to fire. Second, that, when F and G have independent tendencies to cause E, we can easily see that

$$-\log[1 - P(E \mid F \cdot G)] = -\log[1 - P(E \mid F)] - \log[1 - P(E \mid G)] + \log[1 - P(E)]$$

and consequently the strengths would not be additive.

Hopefully, these measures of strengths of association between assemblies will help to suggest some quantitative neural mechanisms that could be put to experimental test.

In physical terms, the interaction between a pair of assemblies, A and B, will depend on the number and size of the subassemblies (including the subsubassemblies) that they have in common. This set of subassemblies could be called the "intersection," $A.B$. (A more complete notation would be $A.B(T)$, where T is the time since B fired. The intersection decreases to zero as T increases.) The second-order interaction between three assemblies, A, B, and C, will depend on the set of subassemblies common to all of them, $A.B.C$. If B and C have just been active, they will contribute a "force" tending to activate A, expressible in the form $|\, A.B \,| + |\, A.C \,| - |\, A.B.C \,|$, where the moduli signs represent in some sense the current total strengths of the sets of subassemblies. The term $|\, A.B.C \,|$ is subtracted in order that it should not be counted twice. More generally, as in the Boole-Poincaré theorem, the firing of an assembly sequence, $A_1, A_2, \ldots A_m$ will have an "intrinsic" tendency to cause A to fire, measured (compare the Appendix) by

$$|\, A \cup (A_1 \cdot A_2 \cdot \ldots \cdot A_m) \,| = \sum_r |\, A \cdot A_r \,| - \sum_{r,s} |\, A \cdot A_r \cdot A_s \,| + \sum_{r,s,t} |\, A \cdot A_r \cdot A_s \cdot A_t \,| - \ldots$$

$(r < s < t < \ldots)$. To this must be added a term depending on the current "force" on A from the centrencephalic system, which will perhaps be a function only of the probability that A fires conditional only on past history and psychological "set." The assembly, A, for which the total causal force is a maximum is the one most likely to fire, or, on a deterministic theory, the one that actually will fire. The formula can be interpreted in various ways, depending on whether we have in mind a theory of primed neurons, a theory of subassemblies, or a mixture of the two if we use the anti-Ockham principle for very complex systems.

We shall now consider another semiquantitative aspect of the interaction between assemblies.

Suppose that A and B are two assemblies having no previous association, but that A happens to occur before B, owing to the sequence of events at the sensorium. Suppose that each of the assemblies contains about a fraction α of the cortex (or of the association areas), where α might be, say, 1/30, although this is in large part a guess, as we said before. The neurons in common will constitute about α^2 of the cortex. The synapses connecting these will undergo a slight change of state, encouraging interfacilitation. Thus the common neurons will have some tendency to include a set of subassemblies containing less than α^2 of the cortex. It is not necessary to assume that the temporal order of A and B

is also represented in the interfacilitation in order that a record be made of the temporal sequence of events, provided that we allow for assembly sequences consisting of more than two assemblies. When we recall some event having extension in time, we need to regenerate an assembly sequence. That this is possible is not surprising in view of the subassembly theory. For each assembly was originally fired by the subassemblies left behind by the previous assemblies of the sequence, so if we have succeeded in recalling most of these assemblies it is likely to be easy to recall the next one (since we shall have injected just about the right collection of subassemblies into our cortex). The subassemblies left in the wake of an assembly sequence $A_1, A_2, \ldots, A_m$ will tend to fire A_{m+1}, not A_o, that is, there will be little tendency to remember serial events in reverse time order.

If assemblies $A_1, A_2, \ldots, A_k$, having no previous association, happen to occur in sequence, where k is not more than about 20, then primitive subassemblies (or classes of subassemblies) (A_1, A_2), $(A_2, A_3), \ldots,$ (A_{k-1}, A_k) will form, and perhaps also some weaker subassemblies (A_r, A_s), where $r < s - 1$. These will be at least analogous to the mutual informations $I(A_r, A_s)$, which, for nonperiodic Markov processes, do tend to be weaker and weaker, the larger is $s - r$. Similarly sets of subassemblies and perhaps subsubassemblies will form, corresponding to triples of assemblies, and analogous to the mutual informations $I(A_r, A_s, A_t)$, and so on, for interactions of higher order. (Similar comments, both here and later, can be made if the strengths of association are defined in terms of K in place of I.) The set of subassemblies arising from the "intersection" of q assemblies of which none had been previously associated, could hardly occupy a proportion of the cortex larger than α^q, so that, if $\alpha = 1/30$, q could not be larger than $\log_{30}(5 \times 10^9) = 6\frac{1}{2}$. This would not constitute a serious biological disadvantage, since high-order interactions can generally be ignored, judging by the practice of statisticians in factorial experiments (see the Appendix). The upper limit is reminiscent of the "depth hypothesis" [*69, 107*]. Compare also the experiment mentioned at the beginning of Section 5.

We have seen that it is impracticable to take a sample of language that is large enough to be able to judge the association factors (the exponentials of the amounts of mutual information) between all pairs of 30,000 words by simple frequency counts. It is reasonable to assume that direct psychological association between words is determined by the frequencies with which they occur nearly simultaneously in thought, and this is easy to understand in a general way in terms of the assembly and subassembly theory. But we can recognize logical associations between pairs of words that have never occurred together in our

experience; for example, the words "ferry" and "fare" can be seen to be associated in the same manner as "bus" and "fare," even if we never previously made the association in our minds. Likewise, if we were asked to estimate the mutual information between the first two words "ferry" and "fare," regarded as index terms for sentences, we could reasonably take it as equal to that between the second pair. This is a simple example to show that we make use of semantics even in the simplest problems of association whenever our samples have not been large enough to rely on mere frequencies. The simplest conditional probability machines, such as those designed by Uttley [*101*], rely only on frequencies, in other words the probabilities are maximum-likelihood estimates, and they make no use of semantics. Such machines could be improved in principle by means of automatic classification of words into "clumps" (see Section 5). The essential idea is that words can be seen to be associated not merely because they occur frequently together, but because they both occur frequently in conjunction with a third word, or more generally with other words that belong to some reasonably objectively definable clump of words. The search for clumps is especially interesting for the purpose of trying to construct a thesaurus mechanically, hopefully for application to problems of classification and mechanical translation. A comprehensive search is liable to be very expensive in computer time, if the computer is of classical design. By using an artificial neural net, it might be possible to perform the search faster, owing to the parallel working. If $A_1, A_2, \ldots, A_k$ is a clump of assemblies having respectively $n_1, n_2, \ldots n_k$ subassemblies, and if A_i and A_j have m_{ij} subassemblies in common; then, for each i, the "clumpiness"

$$\frac{1}{k-1}\sum_j \frac{m_{ij}}{n_i}$$

is much larger than it would be for a random class of k assemblies. One can define a clump by insisting that the clumpiness is decreased if any assembly is added to the clump or removed from it. Many other definitions of a clump are possible (see for example Section 5, and [*31, 41*], and references given in the latter article), and it is not yet clear to what extent the definitions agree with each other, nor which definitions are appropriate for various purposes. At any rate *we must suppose that there is some mechanism by which an assembly representing a clump of assemblies tends to be formed*, a mechanism that will correspond at least to some aspects of "abstraction" or "generalization." Often this assembly will itself represent a word, and the existence of the word will encourage the assembly to form (for example [*41*], p. 122): in the example of ferries and buses the word might be "vehicle." In the design

of an ultraintelligent machine based on an artificial neural net, one of the most vital problems is how to ensure that the above mechanism will be effective. It seems to be necessary to assume that, when an assembly is active, it causes a little activity in all the assemblies with which it is closely associated, although only one at most of these assemblies will be the next to fire. This "priming" of assemblies is analogous to the priming of neurons; it is presumably operated by the subassemblies. The slightly active assemblies in their turn might encourage an even smaller amount of activity in those with which *they* are closely associated. In this way, there will be a small amount of activity in all the assemblies of a clump, although none of them is actually fired, and consequently a gradually increased chance that an assembly will form that will represent a clump. In terms of man, since, by hypothesis, we are not conscious of cortical activity that is not part of an active assembly, when we form a new abstraction it will emerge from the preconscious or unconscious in a manner that will seem to our conscious minds like a flash of inspiration!

It is possible that one of the functions of sleep is to give the brain an opportunity of consolidating the waking experiences by means of unconscious botryological calculations, especially those leading to improved judgments of probabilities. This assumption would be consistent with the advice to "sleep on a problem." It might turn out that an ultraintelligent machine also would benefit from periods of comparative rest, but not by being switched off.

Some of the matters that have been discussed in this section can be apprehended as a whole in terms of the following survey of short-term and long-term memory. In most modern computers there are several levels of storage, successively larger but slower. The reason for this is that it would be too expensive to have an exceedingly large storage with instant recall. It is natural to suppose that human memory too is split up into levels corresponding to different mechanisms. The following classification would be consistent with the discussion in this section. It is of course conjectural.

(i) *Immediate recall (about $\frac{1}{2}$ second)*. Concepts currently in consciousness, embodied in the currently active assembly.

(ii) *Very short-term memory or attention span ($\frac{1}{2}$ second to 10 seconds)*. Embodied in the currently active subassemblies, largely the residues of recently active assemblies. The span might be extended up to several minutes, with embodiment in subsubassemblies, etc.

(iii) *Short-term (from about 10 seconds or 10 minutes to about one day)*. Embodied in primed neurons.

(iv) *Medium-term (about one day to about one month, below the age of 30, or about one week above the age of 50)*. Assemblies are neither partly active nor partly primed, but present only by virtue of their patterns of synaptic strengths, and with little degradation.

(v) *Long-term (about one month to a hundred years)*. As in (iv) but with more degradation of pattern and loss of detail.

A program of research for quantitative theory would be to marry the histological parameters to those in the above list. This program will not be attempted here, but, as promised earlier, we shall give one example of how a quantitative theory might be developed (see also, for example, [*6, 92*]). Let us make the following provisional and artificial assumptions:

(i) The probability, in a new brain, that a pair of neurons is connected is the same for every pair of neurons.

(ii) Each neuron has μ inhibitory synapses on it, and vastly more excitatory ones.

(iii) A single "pulsed" inhibitory synapse dominates any number of pulsed excitatory ones, during a summation interval.

(iv) An assembly occupies a proportion α of the cortex and the active subassemblies not in this assembly occupy a proportion $\beta - \alpha$, making a total activity equal to β.

Then a random neuron has probability $(1 - \beta)^\mu$ of escaping inhibition. In order to be active, the neuron must also escape inhibition by the centrencephalic system. So

$$\beta < (1 - \beta)^\mu$$

Therefore

$$\mu < \frac{\log \beta}{\log(1 - \beta)}$$

For example, if $\beta = 1/15$, then $\mu < 52$. It seems unlikely that any biochemical mechanism could be accurate enough to give the required value of μ, without some feedback control in the maturation of the brain. But it is perhaps significant that the number of neurons in the cortex is about 2^{32}, so that, perhaps, in the growth of the brain, each neuron acquires one inhibitory synapse per generation, 31 in all. The conjecture would have the implication that close neurons would tend to inhibit each other more than distant ones, as required by Milner [*71*] (compare [*34*]). We emphasize that this example is intended only to be illustrative of how a quantitative theory might proceed. Taken at its face value, the example is very much more speculative than the subassembly theory as a whole.

We conclude this section with a brief discussion of an objection that has been made to the assembly theory. Allport ([3], p. 179) says,

regarding the observation of a whole that consists of parts, a, b, c, "... There is, in Hebb's scheme, no apparent reason why we should not have ... a perception of the parts a, b, c and *alongside* these at the same time another equally vivid perception of the whole, that is, of t. This, however, does not occur: we perceive either the parts in their separateness *or* the parts as integrated into a whole, but not both at once" (see Hebb [*54*], pp. 98–99).

This does not seem to be an objection to the theory in the form presented here. Even if the assembly t were originally built up largely from parts of the assemblies a, b, c, it does not contain the whole of any one of these three assemblies. Instead, it consists of parts of a, b, c and also of parts not in a, b, or c. Consequently it is only to be expected that we do not apprehend an object both as a whole and in its parts at quite the same moment.

In the next section we suggest how meaning might be represented in terms of subassemblies, but only in a general manner, and not with the degree of precision that could be desired. We aim mainly to strengthen the case that semantics are relevant to artificial intelligence, and to lend support to the feeling, that is very much in the air at present, that much more detailed research into these matters is worthwhile.

7. An Assembly Theory of Meaning

Our purpose is not to define "meaning," but to consider its physical embodiment. We have already discussed various aspects of meaning in previous sections, and this will enable us to keep the present section, and the next one, short.

A distinction can be made between the literal meaning of a statement, and the subjective meaning that the statement has (on a particular occasion) to a man or machine. It is the latter that is of main concern to us in this essay. (For a man, subjective meaning could also be aptly called "personal meaning" but this name would at present seem inappropriate for a machine.) Although we are concerned with subjective meaning, the behavioral interpretation of meaning is not enough for us, as was said in Section 4. Instead, the subjective meaning of a statement might be interpreted as the set of tendencies to cause the activation of each assembly sequence at each possible time in the future. The physical embodiment of meaning, when a statement is recalled to mind, would then be a class of subassemblies.

This embodiment of meaning is related to the probabilistic interpretation for the meaning of a word, given in Section 4 (the probabilistic form of "Wisdom's cow"). The qualities $Q_1, Q_2, \ldots, Q_m$, when noticed one at a time, would activate assemblies, but, when they are

noticed only preconsciously, and so directly cause activity only in subassemblies, they are at best contributory causal agents in the activation of assemblies.

If a statement provoked an assembly sequence, S_0, presumably the (subjective) meaning of the statement is embodied in some of the subassemblies that were left behind by S_0, the ones that reverberated the longest being the most important ones. Two statements have close meanings if the sets of subassemblies left behind by them bear a close resemblance to each other, or even if the resemblance is not close provided that the *effects* are similar, just as a cow can be recognized on different occasions by the apprehension of different sets of probable properties. We feel that we have understood the meaning of a statement when we somehow recognize that the statement was a definite causal agent in our thought processes or in our propensities to future motor activity, and that these propensities are of a kind which we think was intended by the person who communicated the statement. But I shall ignore these intentions and interpret "meaning" as "meaning for us." Degrees of meaning exist, and correspond in part to greater or lesser degrees of causal tendency.

The "circularity" mentioned in Section 4, in connection with the probabilistic interpretation of meaning, corresponds to the obvious possibility that an assembly can help to strengthen some of the weak subassemblies that helped to activate the assembly itself.

A more formal suggestion for the representation of meaning can be framed as follows.

Let S be an assembly sequence, and $\mathfrak{S}$ a "set" in the psychological sense. (An assembly theory of psychological set is given in Hebb [*54*].) Let Σ be a statement. Denote by

$$P(A \mid S \cdot \mathfrak{S} \cdot \Sigma)$$

the probability that A will be the next dominant assembly to follow the assembly sequence S when the subject is in psychological set $\mathfrak{S}$, and when he has been told Σ and had no reason to doubt the veracity of his informant. If the subject had not been told Σ the corresponding probability would be

$$P(A \mid S \cdot \mathfrak{S})$$

and, if he had been told that Σ was false, the probability would be denoted by

$$P(A \mid S \cdot \mathfrak{S} \cdot \bar{\Sigma})$$

Then the function of A, S, and $\mathfrak{S}$, with values

$$\log[P(A \mid S \cdot \mathfrak{S} \cdot \Sigma)/P(A \mid S \cdot \mathfrak{S} \cdot \bar{\Sigma})] \qquad (7.1)$$

for all A, S, and $\mathfrak{S}$, is a reasonable first approximation to a representation of the "meaning" of Σ. The representation of the meaning of the negation of Σ is minus that of Σ. A reasonable representation of the "effectiveness" of the statement would be the function with values

$$\log[P(A \mid S \cdot \mathfrak{S} \cdot \Sigma)/P(A \mid S \cdot \mathfrak{S})] \tag{7.2}$$

The reason why this latter formula would be inappropriate as a representation of "meaning" is that it is sensitive to the subject's degree of belief in Σ before he is told Σ. A man's degree of belief in a statement should not be very relevant to its meaning.

It is not intended to be implied by this representation that the subject could obtain the values of the probabilities by introspection. The probabilities are intended to be physical probabilities, not the subjective probabilities of the man or machine. (For a discussion of kinds of probability, see, for example, [*36*].)

Expression (7.1) may be described as the log-factor or weight of evidence in favor of the hypothesis that Σ was stated rather than $\bar{\Sigma}$, provided by the event that assembly A was activated, given that the previous assembly sequence was S, and that the psychological set was $\mathfrak{S}$. (The terminology is that of [*26*] and [*21*], for example, and was mentioned in Section 5.) If the subject is deterministic, then the probabilities would be pseudoprobabilities, of the same logical nature as those associated with pseudorandom numbers. Expression (7.2) is the mutual information between the propositions that the assembly A was activated on the one hand and that Σ was stated on the other.

If the class of values of (7.1) is extended also over several subjects (who could be specified in the notation) then we should have a representation of multisubjective meaning, and we might perhaps approximate to a representation of "true meaning" if there is such a thing. A representation of "literal meaning" could be obtained by restricting the class to "literal-minded" men and robots, in order to exclude the poetic and irrational influences of a statement.

Formulas (7.1) and (7.2) are of course only examples of possible quantitative representations of "meaning." It might be better to replace them by the formulas for causal tendency.

$$Q(A : \Sigma \mid S \cdot \mathfrak{S}) = \log \frac{1 - P(A \mid S \cdot \mathfrak{S} \cdot \bar{\Sigma})}{1 - P(A \mid S \cdot \mathfrak{S} \cdot \Sigma)} \tag{7.1a}$$

and

$$K(A : \Sigma \mid S \cdot \mathfrak{S}) = \log \frac{1 - P(A \mid S \cdot \mathfrak{S})}{1 - P(A \mid S \cdot \mathfrak{S} \cdot \Sigma)} \tag{7.1b}$$

These formulas would be more consistent with the interpretation of the meaning of a statement in terms of its causal propensities.

Although we are arguing that semantics are relevant to the design of an ultraintelligent machine, we consider that it will not be necessary to solve all of the problems of semantics in order to construct the machine. If we were using the approach depending on a "canonical language" (see Section 4), the problems would all need solution, but if a neural net is used, we believe that the net might be capable in effect of learning semantics by means of positive and negative reinforcement, in much the same manner as a child learns. The theory of assemblies and subassemblies, as applied to semantics, is intended to provide some at least intuitive justification for this belief. It should be possible, by means of more quantitative theory and experiment, to improve, to disprove, or to prove the theory. A thoroughgoing quantitative theory will be difficult to formulate, and the experiments will be laborious and expensive, but the reward or punishment will be great.

8. The Economy of Meaning

Just as the activation of an assembly is a form of regeneration, so also is that of a subassembly, although the regeneration of subassemblies might be less sharp. The degree of regeneration of a subassembly corresponds to a preconscious estimate of the probability of some property, so that the process of recall is physically one of regeneration mixed with probabilistic regeneration. We have argued that, in any communication system, the function of regeneration and of probabilistic regeneration is economy, and so the physical embodiment of meaning also serves a function of economy. It is even possible that the evolutionary function of meaning and understanding is economy, although metaphysically we might consider that the function of evolution is the attainment of understanding!

Imagine, for the sake of argument, that each meaningful proposition (defined as a class of logically equivalent statements) could be expressed by each of a hundred different statements, each of which had an entirely distinct representation in the brain. Suppose that the number of ordered pairs of propositions that are mentally associated is N. Corresponding to each pair of propositions, there would be 10,000 equivalent pairs of statements. In order to represent the N associations between propositions, we should require $10{,}000N$ associations between statements. Although the number 100 is here a pure guess, it is clear that there must be a tremendous premium on the representation of statements by their meanings. For this saves a factor of 100 (nominally) in the storage of the propositions, and a corresponding factor of 10,000 in the storage of the associations between pairs of propositions. The latter factor is relevant in *long-term recall*, since the process of recalling

a fact usually requires that one should have in mind several other facts. It is clear therefore that the physical representation of meaning performs a very important function of economy, especially in long-term recall, and can be expected to perform an equally important function in an ultraintelligent machine.

9. Conclusions

These "conclusions" are primarily the opinions of the writer, as they must be in a paper on ultraintelligent machines written at the present time. *In the writer's opinion then*:

It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make, since it will lead to an "intelligence explosion." This will transform society in an unimaginable way. The first ultraintelligent machine will need to be ultraparallel, and is likely to be achieved with the help of a very large artificial neural net. The required high degree of connectivity might be attained with the help of microminiature radio transmitters and receivers. The machine will have a multimillion dollar computer and information-retrieval system under its direct control. The design of the machine will be partly suggested by analogy with several aspects of the human brain and intellect. In particular, the machine will have high linguistic ability and will be able to operate with the meanings of propositions, because to do so will lead to a necessary economy, just as it does in man.

The physical representation of both meaning and recall, in the human brain, can be to some extent understood in terms of a subassembly theory, this being a modification of Hebb's cell assembly theory. A similar representation could be used in an ultraintelligent machine, and is a promising approach.

The subassembly theory leads to reasonable and interesting explanations of a variety of psychological effects. We do not attempt to summarize these here, but merely refer the reader back to Section 6. Even if the first ultraintelligent machine does not after all incorporate a vast artificial neural network, it is hoped that the discussion of the subassembly theory is a contribution to psychology, and to its relationships with the theories of communication and causality.

The activation of an assembly or a subassembly is an example of generalized regeneration, a function of which is again economy. The assembly and subassembly theories are easier to accept if combined with the assumption of a centrencephalic control system, largely because this would enable a very much greater variety of assemblies to exist.

The process of long-term recall can be partly understood as a statistical information-retrieval system. Such a system requires the estimation

of probabilities of events that have never occurred. The estimation of such probabilities requires some nontrivial theory even in simple cases, such as for multinomial distributions having a large number of categories. In more complicated cases, the theories are very incomplete, but will probably require a knowledge of and an elaboration of all the methods that have so far been used by actuaries and other statisticians for the estimation of probabilities. Among the techniques will be included the maximum-entropy principle, the use of initial probability distributions [*47*, *56*, *48*], and "botryology" (the theory and practice of clump-finding).

A form of Bayes' theorem expresses the final log-probability of a "document" or "memory" as an initial log-probability, plus some terms representing $I(D_j : W_i)$, the information concerning a document provided by an index term (or concerning a memory provided by a "clue"), plus additional terms representing the mutual information between index terms and the document. It is suggested that, in the brain, the initial log-probability is possibly represented in some sense by the strength of the connectivity between an assembly and the centrencephalic system; that the terms $I(D_j : W_i)$ are represented by the subassemblies shared between the assemblies corresponding to D_j and W_i; and that other terms are represented by the interactions between sets of at least three assemblies.

An alternative suggestion, which seems slightly to be preferred, is that the strengths of association are expressible in terms of $K(E : F)$, the intrinsic tendency of an event E to be caused by F. This is equal to minus the mutual information between F and not E. Then the strength of the association from the centrencephalic system and an assembly would be approximately equal to the initial (prior) probability of the firing of the assembly, given the psychological "set." The same remarks concerning interactions apply here as in the first suggestion.

Whereas, in ordinary information-retrieval problems, the expression $I(D_j : W_i)$ will often need to be estimated with the help of computational techniques for clumping, the strength of the connectivity between two assemblies will often be physically represented because of the manner in which the two assemblies were originally formed, by being built up from co-occurring subassemblies.

The representation of informational or causal interactions, or both, up to about the sixth or seventh order, is presumably embodied in the subassemblies common to assemblies. The magical proficiency of the brain, in recall, can be largely attributed to its facility in handling these interactions. My guess is that only an ultraparallel machine, containing millions of units capable of parallel operation, could hope to compete with the brain in this respect.

It seems reasonable to conjecture that the organization of the interactions into subassemblies might require the intervention of periods of rest or sleep. A possible function of sleep is to replay the assembly sequences that were of greatest interest during the day in order to consolidate them. During wakefulness, half-formed subassemblies would be subjected to the inhibitory effect of fully active assemblies, but during sleep a half-formed subassembly would have time to organize and consolidate itself. On this hypothesis, a function of sleep is to strengthen the unconscious and preconscious parts of the mind.

The first ultraintelligent machine will be educated partly by means of positive and negative reinforcement. The task of education will be eased if the machine is somewhat of a robot, since the activity of a robot is concrete.

Regarding the microstructure of the learning process, it is proposed that this be effected by means of reinforcement of the strengths of artificial synapses, that the available strengths for each synapse should form a discrete set, that when a synapse is not used for a certain length of time it should have a certain small probability of "mutation" down one step, and that when a synapse is "successfully used" (i.e., contributes to the activation or inhibition of an artificial neuron) it has a certain small probability of mutation up one step. The need for the changes in synaptic strength to be only probabilistic, with small probabilities, is that they would otherwise vary too quickly for the machine to be of any use, at any rate if the assembly or subassembly theory is incorporated. Deterministic changes, in any obvious sense, would be useful only if a very small fraction of the machine were in use at one instant, and this would be uneconomical.

10. Appendix: Informational and Causal Interactions

Let $E_1, E_2, \ldots, E_n$ represent events or propositions. Let the probability $P(E_1 \cdot \bar{E}_2 \cdot \ldots \cdot \bar{E}_n)$, for example, where the vinculum denotes negation, be denoted by $p_{10\cdots1}$, where 0 means false and 1 means true. The 2^n different possible logical conjunctions of the n propositions and their negations have probabilities denoted by p_i, where $i = (i_1, i_2, \ldots, i_n)$ is an n-dimensional vector each of whose components is either 0 or 1. The array (p_i) is a 2^n *population contingency table.*

A *marginal total* of the table is obtained by summing out one or more of the suffixes, and we denote $\Sigma_{i_3,i_5} p_i$, for example, by $p_{i_1 i_2 \prime i_4 \prime i_6 i_7 \ldots i_n}$. When the suffixes not summed out are equal to 1, we use an alternative notation: for example, if $i_1 = i_2 = i_4 = i_6 = i_7 = \ldots = i_m = 1$, we denote the marginal total by $P_{11010111\ldots1}$. Thus the numbers (P_i) form

another 2^n array, which consists of a subset of the marginal totals of the original table. Note that $P_{000\ldots0} = p_{\prime\prime\prime\ldots\prime} = 1$, and that, for example, $P_{11000\ldots0} = P(E_1 \cdot E_2)$. The probabilities (P_i) have more direct relevance to information retrieval than the probabilities (p_i), since it is more natural to assert an index term than to deny one. The most relevant P_i's for this purpose will be those for which $|\, i \,|$ is small, where $|\, i \,|$ denotes the number of nonzero components of i.

Just as each P_i is the sum of some of the p_i's, so each p_i can be expressed as a linear combination of P_i's. For example (the Boole-Poincaré theorem),

$$\begin{aligned} p_{000\ldots0} &= P(\bar{E}_1 \cdot \bar{E}_2 \cdot \ldots \cdot \bar{E}_n) \\ &= \sum (-1)^{|\,i\,|} P_i \\ &= S_0 - S_1 + S_2 - \ldots + (-1)^n S_n \end{aligned} \tag{A.1}$$

where S_μ is the sum of all P_i's for which $|\, i \,| = \mu$, and

$$p_{1000\ldots0} = S_1' - S_2' + S_3' - \ldots$$

where S_μ' is the sum of all P_i's for which $|\, i \,| = \mu$ and $i_1 = 1$.

Interactions between events: Let E and F be two propositions or events. Write

$$I(E) = -\log P(E)$$

the amount of information in the proposition E concerning itself ([*21*], p. 74, [*26*]). Let

$$\begin{aligned} I(E : F) &= I(E, F) = \log P(E \cdot F) - \log P(E) - \log P(F) \\ &= I(E) + I(F) - I(E \cdot F) \end{aligned} \tag{A.2}$$

the *amount of information concerning E provided by F*. It is also called the *mutual information* between E and F, when it is desired to emphasize the symmetry, and in this case the comma is more appropriate than the colon, since the colon is pronounced "provided by." The equation shows that $I(E, F)$ can be regarded as a measure of *information interaction* between E and F. For sets of more than two propositions, it is natural to generalize this definition by using the n-dimensional mod 2 discrete Fourier transform, as in the definition of interactions for factorial experiments (see for example [*32*]). We write

$$I_j = 2^{\,|\,j\,|\,-\,n} \sum_i (-1)^{ij} \log P_{i \wedge j} \tag{A.3}$$

where $ij = i_1 j_1 + \ldots + i_n j_n$ is the inner product of i and j, and $i \wedge j = (i_1 j_1, \ldots, i_n j_n)$ is the "indirect product" (halfway between the inner

product and the direct product). We call I_j an (*informational*) *interaction of the first kind and order* $|j|-1$. For example,

$$
\begin{aligned}
I_{000\ldots0} &= \log P_{000\ldots0} = 0 \\
I_{100\ldots0} &= \log P_{000\ldots0} - \log P_{100\ldots0} = -\log P(E_1) = I(E_1) \\
I_{110\ldots0} &= I(E_1) + I(E_2) - I(E_1 \cdot E_2) = I(E_1, E_2) \\
I_{1110\ldots0} &= I(E_1) + I(E_2) + I(E_3) - I(E_2 \cdot E_3) \\
&\quad - I(E_3 \cdot E_1) - I(E_1 \cdot E_2) + I(E_1 \cdot E_2 \cdot E_3) \\
&= I(E_3 : E_1) + I(E_3 : E_2) - I(E_3 : E_1 \cdot E_2) \qquad \text{(A.4)} \\
I_{11110\ldots0} &= I(E_4 : E_1) + I(E_4 : E_2) + I(E_4 : E_3) - I(E_4 : E_2 \cdot E_3) \\
&\quad - I(E_4 : E_1 \cdot E_3) - I(E_4 : E_1 \cdot E_2) + I(E_4 : E_1 \cdot E_2 \cdot E_3) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(A.5)}
\end{aligned}
$$

In [*37*], this last expression was denoted by $I_2(E_4 : E_1 \cdot E_2 \cdot E_3)$, but $I_3(E_4 : E_1 \cdot E_2 \cdot E_3)$ would be a more natural notation. We avoid this notation here since we are using vectors for suffixes. We prefer to write $I_{11110\ldots0} = I(E_1, E_2, E_3, E_4)$, and regard it as the mutual information between the four propositions. By means of the Fourier inversion formula, we see, for example, that (as in [*16a*], p. 58)

$$
\begin{aligned}
-\log P_{11110\ldots0} &= I(E_1 \cdot E_2 \cdot E_3 \cdot E_4) \\
&= \sum I(E_r) - \sum I(E_r, E_s) \\
&\quad + \sum I(E_r, E_s, E_t) - I(E_1, E_2, E_3, E_4) \qquad \text{(A.6)}
\end{aligned}
$$

where $1 \leq r < s < t \leq 4$. Equation (5.6) is readily deducible.

Interactions between causal tendencies (Q or K, see page 67), are definable in a similar manner. (Compare [*39*], where the signs are not quite the same as here.) But we shall leave these to the reader's imagination.

We also write

$$J_j = \sum_i (-1)^{ij} \log p_i \qquad \text{(A.7)}$$

and call J_j an (*informational*) *interaction of the second kind, of order* $|j|-1$. (It was denoted by I_j in [*47*].) Yet another kind of interaction, involving expected amounts of information, was defined by McGill and Quastler [*67*].

If n is not small, the number of cells in the population contingency table is very large, and an exceedingly large sample would be required in order to make a direct estimate of all the p_i's. In order to get around this difficulty to some extent we can sample just *some* of the marginal totals. Then we can use the "maximum-entropy principle" [*47, 48, 52, 55, 56, 62, 100*] in order to make at least a provisional estimate of the p_i's. According to this principle, one maximizes the entropy $-\sum p_i \log p_i$, subject to the constraints (here, the assigned marginal totals) in order to set up a null hypothesis (at least this is the way it is expressed in [*47*] and [*48*]). The idea in a nutshell is to assume as much statistical independence as one can. Among other things, the following result is proved in [*47*]:

Suppose that we know or assume a complete set of rth-order constraints for (p_i), i.e., all totals of the p_i's over each subset of $n - r$ coordinates. Then the null hypothesis generated by the principle of maximum entropy is the vanishing of all the rth and higher-order interactions of the second kind.

In this theorem, instead of assuming a complete set of rth-order constraints, we could assume all the interactions of the first kind and orders $r-1$ or less.

In order to see this, we take $r = 4$ for simplicity and consider Eq. (A.3). If we know P_i for all i with $|i| \leq 4$, we can calculate all I_j with $|j| \leq 4$, i.e., we can deduce all the interactions of the first kind and of orders 3 or less. Conversely, given these interactions of the first kind, we can first calculate $\log P_{10000}$, $\log P_{0100}$, $\log P_{0010}$, $\log P_{0001}$, then $\log P_{1100}$ (since we know $\log P_{1100} - \log P_{1000} - \log P_{0100}$), and so on. We can thus determine P_i for all i with $|i| \leq 4$, i.e., we have a complete set of fourth-order constraints of the p_i's.

Nearly always, when a statistician discusses interactions of any kind, he believes or hopes that the high-order interactions will be negligible. The maximum-entropy principle provides a comparatively new kind of rationale for this belief regarding interactions of the second kind. Whether a similar partial justification can be provided for other kinds of interaction is a question that has not yet been investigated. The question is analogous to that of the truncation of power series and series of orthogonal functions, as in polynomial approximation.

References[4]

1. *Artificial Intelligence*. IEEE Publ. No. S-142, New York (January 1963).
2. Adrian, E. D., and Matthews, B. H. C., The interpretation of potential waves in the cortex. *J. Physiol. (London)* **81**, 440–471 (1934).
3. Allport, F. H., *Theories of Perception and the Concept of Structure*, Chapter 19. Wiley, New York, 1955.
4. Ashby, W. R., *Design for a Brain*. Chapman & Hall, London, 1960.
5. Bar-Hillel, Y., Semantic information and its measures, *Trans. 10th Conf. Cybernetics, New York*, pp. 33–48 (1953).

[4]Only items mentioned in the text are listed here; for a bibliography on artificial intelligence see Minsky [*72*].

6. Beurle, R. L., Functional organization in random networks. In *Principles of Self-Organization* (H. von Foerster and G. W. Zopf, Jr., eds.), pp. 291–311 and discussion pp. 311–314. Oxford Univ. Press, London and New York, 1962.
7. Black, M., *Language and Philosophy*. Cornell Univ. Press, Ithaca, New York, 1949.
8. Brain, Lord, Recent work on the physiological basis of speech. *Advancement Sci.* **19**, 207–212 (1962).
9. Bruner, J. S., Postman, L., and Rodrigues, J., Expectation and the perception of color. *Am. J. Psychol.* **64**, 216–227 (1951).
10. Carnap, R., and Bar-Hillel, Y., Semantic information. *Brit. J. Phil. Sci.* **4**, 147–157 (1953).
11. Carter, C. F., Problems of economic growth. *Advancement Sci.* **20**, 290–296 (1963).
12. Cherry, E. C. *On Human Communication*. Wiley, New York, 1957.
13. Cherry, E. C., Two ears—but one world. In *Sensory Communication* (W. A. Rosenblith, ed.), pp. 99–116. Wiley, New York, 1961.
14. Chomsky, N., Explanatory models in linguistics. In *Logic, Methodology and Philosophy of Science* (E. Nagel, P. Suppes, and A. Tarski, eds.), pp. 528–550. Stanford Univ. Press, Stanford, California, 1962.
15. Culbertson, J. T., *Consciousness and Behavior*. W. C. Brown. Dubuque, Iowa, 1950.
16. Eccles, J. C., *Physiology of Nerve Cells*. Johns Hopkins Press, Baltimore, Maryland, 1957.

16a. Fano, R. M., *Transmission of Information*. Wiley, New York, 1961.

17. Flory, P. J., *Principles of Polymer Chemistry*. Cornell Univ. Press, Ithaca, New York, 1953.
18. Friedberg, R. M., A learning machine, Part I. *IBM J. Res. Develop.* **2**, 2–13 (1958).
19. Gabor, D., Wilby, W. P. L., and Woodcock, R., A self-optimizing nonlinear filter, predictor and simulator. In *Information Theory: Fourth London Symposium* (E. C. Cherry, ed.), pp. 348–352. Butterworth, London and Washington, D.C., 1961.
20. Gall, F. J., and Spurzheim, G., *Anatomie et Physiologie du Système Nerveux en Général et du Cerveau en Particulier, avec des Observations sur la Possibilité de Reconnaître Plusiuers Dispositions Intellectuelles et Morales de l'Homme et des Animaux par la Configuration de Leurs Têtes*, 4 vols. Paris, 1810–1819. (Cited in Zangwill [*108*].)Volumes 3 and 4 are by Gall alone.
21. Good, I. J., *Probability and the Weighing of Evidence*. Hafner, New York, 1950.
22. Good, I. J., Review of a book by D. R. Hartree. *J. Roy. Statist. Soc.* **A114**, 107 (1951).
23. Good, I. J., Rational decisions. *J. Roy. Statist. Soc.* **B14**, 107–114 (1952).
24. Good, I. J., in *Communication Theory* (W. Jackson, ed.), p. 267. Butterworth, London and Washington, D.C., 1953.
25. Good, I. J., On the population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264 (1953).
26. Good, I. J., Some terminology and notation in information theory, *Proc. IEE (London) Part C (3)* **103**, 200–204 (1956).
27. Good, I. J., On the estimation of small frequencies in contingency tables. *J. Roy. Statist. Soc.* **B18**, 113–124 (1956).
28. Good, I. J., Distribution of word frequencies. *Nature* **179**, 595 (1957).

29. Good, I. J., Review of a book by G. S. Brown. *Brit. J. Phil. Sci.* **9**, 254 (1958).
30. Good, I. J., How much science can you have at your fingertips? *IBM J. Res. Develop.* **2**, 282–288 (1958).
31. Good, I. J., Speculations concerning information retrieval. Research Rept. No. RC-78, IBM, Yorktown Heights, New York (1958).
32. Good, I. J., The interaction algorithm and practical fourier analysis, *J. Roy. Statist. Soc.* **B20**, 361–372 (1958); **B22**, 372–375 (1960).
33. Good, I. J., Could a machine make probability judgments? *Computers Automation* **8**, 14–16 and 24–26 (1959).
34. Good, I. J., Speculations on perceptrons and other Automata. Research Rept. No. RC-115, IBM, Yorktown Heights, New York (1959).
35. Good, I. J., *Proc. Intern. Conf. Sci. Inform.*, pp. 1404 and 1406. Natl. Acad. Sci. and Natl. Research Council, Washington, D.C., 1959.
36. Good, I. J., Kinds of probability. *Science* **129**, 443–447 (1959). Italian translation in *L'Industria*, 1959.
37. Good, I. J., Effective sampling rates for signal detection or can the Gaussian model be salvaged? *Inform. Control* **3**, 116–140 (1960).
38. Good, I. J., Weight of evidence, causality, and false-alarm probabilities. In *Information Theory: Fourth London Symposium* (E. C. Cherry, ed.), pp. 125–136. Butterworth, London and Washington, D.C., 1961.
39. Good, I. J., A causal calculus, *Brit. J. Phil. Sci.* **11**, 305–319 (1961); **12**, 43–51 (1961); **13**, 88 (1962).
40. Good, I. J., How rational should a manager be? *Management Sci.* **8**, 383–393 (1962). To be reprinted, with minor corrections, in *Executive Readings in Management Science*, Vol. I (M. K. Starr, ed.). Macmillan, New York, 1965, in press.
41. Good, I. J., Botryological speculations. In *The Scientist Speculates* (I. J. Good, A. J., Mayne, and J. Maynard Smith, eds.) pp. 120–132. Basic Books, New York, 1963.
42. Good, I. J., Review of a book by J. Wolfowitz. *J. Roy. Statist. Soc.* **A125**, 643–645 (1962).
43. Good, I. J., The mind-body problem, or could an android feel pain? In *Theories of the Mind* (J. M. Scher, ed.), pp. 490–518. Glencoe, New York, 1962.
44. Good, I. J., The social implications of artificial intelligence. In *The Scientist Speculates* [*41*], pp. 192–198.
45. Good, I. J., Cascade theory and the molecular weight averages of the Sol fraction. *Proc. Roy. Soc.* **A272**, 54–59 (1963).
46. Good, I. J., The relevance of semantics to the economical construction of an artificial intelligence. In *Artificial Intelligence* [*1*], pp. 157–168.
47. Good, I. J., Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34**, 911–934 (1963).
48. Good, I. J., *The Estimation of Probabilities*. M.I.T. Press, Cambridge, Massachusetts, 1965.
49. Good, I. J., The human preserve. *Spaceflight* in press (1965).
50. Greene, P. H., An approach to computers that perceive, learn, and reason. *Proc. Western Joint Computer Conf.*, pp. 181–186 (1959).
51. Halmos, P. R., *Naive Set Theory*. Van Nostrand, Princeton, New Jersey, 1960.
52. Hartmanis, J., The application of some basic inequalities for entropy. *Inform. Control.* **2**, 199–213 (1959).
53. Hayek, F. A., *The Sensory Order*. Univ. of Chicago Press, Chicago, Illinois, 1952.

54. Hebb, D. O., *Organization of Behavior*. Wiley, New York, 1949.
55. Jaynes, E. T., Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957); **108**, 171–190 (1957).
56. Jaynes, E. T., New engineering applications of information theory. *Proc. First-Symp. Eng. Applications of Function Theory and Probability* (J. L. Bogdanoff and F. Kozin, eds.), pp. 163–203. Wiley, New York, 1963.
57. John, E. R., Some speculations on the psychophysiology of mind. In *Theories of the Mind* [*43*], pp. 80–121.
58. Johnson, W. E., Appendix (ed. by R. B. Braithwaite) to Probability: deductive and inductive problems. *Mind* **41**, 421–423 (1932).
59. Kalmus, H., Analogies of language to life. In *The Scientist Speculates* [*41*], pp. 274–279.
60. Kiseda, J. R., Peterson, H. E., Seelbach, W. C., and Teig, M., A magnetic associative memory, *IBM J. Res. Develop.* **5**, 106–121 (1961).
61. Lashley, K. S., In search of the engram. *Symp. Soc. Exptl. Biol.* **4**, 454–482 (1950).
62. Lewis, P. M., II, Approximating probability distributions to reduce storage requirements. *Inform. Control* **2**, 214–225 (1959).
63. MacKay, D. M., The epistomological problem for Automata. In *Automata Studies* (C. E. Shannon and J. McCarthy, eds.), pp. 235–251. Princeton Univ. Press, Princeton, New Jersey, 1956.
64. Maron, M. E., and Kuhns, J. L., On relevance, probabilistic indexing and information retrieval. *J. Assoc. Computing Machinery* **7**, 216–244 (1960).
65. McDermid, W. L., and Peterson, H. E., A magnetic associative memory system, *IBM J. Res. Develop.* **5**, 59–62 (1961).
66. McDougall, W., *Primer of Physiological Psychology*. Dent, London, 1905.
67. McGill, W., and Quastler, H., Standardized nomenclature: an attempt. In *Information Theory in Psychology* (H. Quastler, ed.), pp. 83–92. Glencoe, New York, 1955.
68. Middleton, D., and Van Meter, D., Detection and extraction of signals in noise from the point of view of statistical decision theory. *J. SIAM* **3**, 192–253 (1956); **4**, 86–119 (1956).
69. Miller, G. A., Human memory and the storage of information. *IRE Trans. Information Theory* **2**, 129–137 (1956).
70. Miller, G. A., and Selfridge, J. A., Verbal context and the recall of meaningful material. *Am. J. Psychol.* **63**, 176–185 (1950).
71. Milner, P. M., The cell assembly: Mark II. *Psychol. Rev.* **64**, 242–252 (1957).
72. Minsky, M., A selected descriptor-indexed bibliography to the literature on artificial intelligence. *IRE Trans. Human Factors in Electron.* **2**, 39–55 (1961).
73. Minsky, M., and Selfridge, O. G., Learning in random nets. In *Information Theory: Fourth London Symposium* (E. C. Cherry, ed.), pp. 335–347. Butterworth, London and Washington, D. C., 1961.
74. Mueller, P., Principles of temporal recognition in artificial neuron nets with application to speech recognition. In *Artificial Intelligence* [*1*], pp. 137–144.
75. Needham, R. M., Research on information retrieval, Classification and Grouping. Rept. No. M.-L.-149, Cambridge Language Research Unit, 1961.
76. Needham, R. M., A method for using computers in information classification. In *Information Processing 1962*, (M. Popplewell, ed.), pp. 284–287. North-Holland Publ. Co., Amsterdam, 1963.
77. Neyman, J., and Scott, E. L., Statistical approach to problems of cosmology. *J. Roy. Statist. Soc.* **B20**, 1–43 (1958).

78. Parker-Rhodes, A. F., Notes for a prodromus to the theory of clumps. Rept. No. LRU-911.2, Cambridge Language Research Unit, 1959.
78a. Pask, G., A discussion of artificial intelligence and self-organization. *Advan. Computers* **5**, 109–226 (1964).
79. Penfield, W., and Jasper, H., Highest level seizures. *Res. Publ. Assoc. Nervous Mental Disease* **26**, 252–271 (1947).
80. Penfield, W., and Roberts, L., *Speech and Brain Mechanisms*. Princeton Univ. Press, Princeton, New Jersey, 1959.
80a. Pierce, J. R., *Symbols, Signals and Noise: the Nature and Process of Communication*. Hutchinson, London, 1962.
81. Rao, C. Radhakrishna, *Advanced Statistical Methods in Biometric Research*, pp. 364–378. Wiley, New York, 1950.
82. Rosenblatt, F., *Principles of Neurodynamics* (Cornell Aeron. Lab., 1961). Spartan Books, Washington, D. C., 1962.
83. Samuel, A. L., Some studies in machine learning using the game of chequers. *IBM J. Res. Develop.* **3**, 210–229 (1959).
83a. Samuel, A. L., Programming computers to play games. *Advan. Computers* **1**, 165–192 (1959).
84. Scriven, M., The compleat robot: a prolegomena to androidology. In *Dimensions of Mind* (S. Hook, ed.), pp. 118–142. N.Y.U. Press, New York. 1960.
85. Sebestyen, G. S., *Decision-Making Processes in Pattern Recognition*. Macmillan, New York, 1962.
86. Selfridge, O. G., Pandemonium: a paradigm for learning. In *Mechanization of Thought Processes*, pp. 511–526. H.M.S.O., London, 1959.
87. Serebriakoff, V., A hypothesis of recognition. In *The Scientist Speculates* [*41*], pp. 117–120.
88. Shannon, C. E., Prediction and entropy of printed English, *Bell System Tech. J.* **30**, 50—64 (1951).
89. Shannon, C. E., and Weaver, W., *The Mathematical Theory of Communication*. Univ. of Illinois Press, Urbana, Illinois, 1949.
90. Sholl, D. A., *The Organization of the Cerebral Cortex*, pp. 5 and 35. Wiley, New York, 1956.
91. Shoulders, K. R., Microelectronics using electron-beam-activated machining Techniques. *Advan. Computers* **2**, 135–293 (1961).
92. Smith, D. R., and Davidson, C. H., Maintained activity in neural nets. *J. Assoc. Computing Machinery* **9**, 268–279 (1962).
93. Sneath, P. H. A., Recent developments in theoretical and quantitative taxonomy. *System. Zool.* **10**, 118–137 (1961).
94. Solomon, R. L., and Howes, D. H., Word frequency, personal values, and visual duration thresholds. *Psychol. Rev.* **58**, 256–270 (1951).
95. Spärck Jones, K., Mechanized semantic classification. *1961 Intern. Conf. on Machine Translation of Languages and Applied Language Analysis*, pp. 417–435. National Physical Laboratory, Teddington, England, 1963.
96. Stiles, H. E., Association factor in information retrieval. *J. Assoc. Computing Machinery* **8**, 271–279 (1961).
97. Tanimoto, T. T., An elementary mathematical theory of classification and prediction. IBM, Yorktown Heights (November 1958).
98. Tompkins, C. B., Methods of successive restrictions in computational problems involving discrete variables, Section IV. *Proc. Symp. Appl. Math.* **15**, 95–106 (1963).
99. Tower, D. B., Structural and functional organization of mammalian

cortex: the correlation of neurone density with brain size. *J. Comp. Neurol.* **101**, 19–46 (1954).
100. Tribus, M., Information theory as the basis for thermostatics and thermodynamics. *J. Appl. Mech.* **28**, 1–8 (1961).
101. Uttley, A. M., The design of conditional probability computers. *Inform. Control* **2**, 1–24 (1959).
102. Uttley, A. M., Conditional probability computing in the nervous system. *Mechanization of Thought Processes*. National Physical Laboratory Symp. No. 10, pp. 119–147 (esp. p. 144, with a reference to an unpublished paper by G. Russell). H.M.S.O., London, 1959.
103. Walter, W. G., *The Living Brain*. Norton, New York, 1953.
104. Winder, R. O., Threshold logic in artificial intelligence. In *Artificial Intelligence* [*1*], pp. 107–128.
105. Woodward, P. M., and Davies, I. L., A theory of radar information. *Phil. Mag.* [*7*], **41**, 1001–1071 (1950).
106. Wozencraft, J. M., and Reiffen, B., *Sequential Coding*. Wiley, New York, 1961.
107. Yngve, V. H., The depth hypothesis. In *Structure of Language and its Mathematical Aspects* (R. Jakobson, ed.), pp. 130–138. Am. Math. Soc., Providence, Rhode Island, 1961.
108. Zangwill, O. L., The cerebral localisation of psychological function. *Advancement Sci.* **20**, 335–344 (1963).